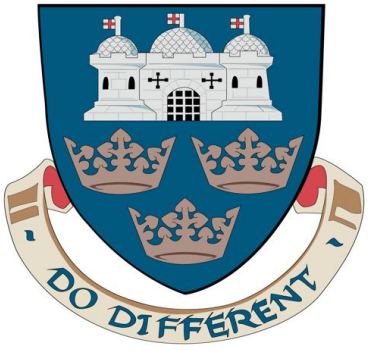
Discovering Dynamic Visemes

Sarah Louise Taylor

A thesis submitted for the Degree of
Doctor of Philosophy

University of East Anglia
School of Computing Sciences



May, 2013

Abstract

This thesis introduces a set of new, dynamic units of visual speech which are learnt using computer vision and machine learning techniques. Rather than clustering phoneme labels as is done traditionally, the visible articulators of a speaker are tracked and automatically segmented into short, visually intuitive speech gestures based on the dynamics of the articulators. The segmented gestures are clustered into *dynamic visemes*, such that *movements* relating to the same visual function appear within the same cluster. Speech animation can then be generated on any facial model by mapping a phoneme sequence to a sequence of dynamic visemes, and stitching together an example of each viseme in the sequence. Dynamic visemes model coarticulation and maintain the dynamics of the original speech, so simple blending at the concatenation boundaries ensures a smooth transition. The efficacy of dynamic visemes for computer animation is formally evaluated both objectively and subjectively, and compared with traditional phoneme to static lip-pose interpolation.

Contents

List of Figures

List of Tables

List of Publications

The following relevant publications were published by the author:

- S. Taylor, B. Theobald, M. Mahler, and I. Matthews (2012). *Dynamic Units of Visual Speech*. In Proceedings of the ACM/ Eurographics Symposium on Computer Animation (SCA), pages 275–284.
- S. Hilder, B. Theobald and R. Harvey (2010). *In Pursuit of Visemes*. In Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP), pages 154–159.
- S. Hilder, R. Harvey and B. Theobald (2009). *Comparison of Human and Machine-based Lip-reading*. In Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP), pages 86–89.

Acknowledgements

First and foremost I would like to thank Dr Barry-John Theobald for proposing such an interesting project, and for his continuous and invaluable support, advice and wisdom. I am also very grateful to Dr Iain Matthews for his significant guidance and encouragement throughout the project, and for allowing me to use his face tracking software. Thank you to Moshe Mahler for providing a 3D face model, for his time spent meticulously animating dynamic visemes on the model and for his help rendering the animated scenes. For help with Maya scripting, I would also like to thank Valeria Reznitskaya. Thank you to all of the participants who took part in the subjective experiments. Finally, thanks to John and the rest of the family for their support and patience.

Chapter 1

Introduction

Realistic facial animation requires care and painstaking manual effort. This is especially true during speech as viewers are extremely sensitive to any discrepancy between sounds and the accompanying facial movements [139]. As the visual quality of computer graphics facial models improve, one also might expect the quality of animated facial behaviour to follow. However, this generally has not been the case and practical applications of speech animation in games and movies is achieved using hand-crafted animation or using expensive motion capture systems.

There are a number of factors that compound the difficulty of synthesising realistic facial movements during speech. Firstly, the biomechanics of the face are complex and it is not clear how these should best be modelled for speech or other facial expressions. Secondly, it is not clear how the underlying visual speech signal should be represented at a segmental level for synthesis. Typically this is done using visemes (visual phonemes [49]) which are traditionally defined as the clusters of visually contrastive phonemes, but whilst deriving a visual unit based on speech acoustics may be convenient as the two modalities are intrinsically linked, this simple approach has a number of problems. The number of phonemes and visemes in an utterance transcription are generally considered to be the same and phoneme labels are simply substituted for viseme labels. Coarticulation effects, where neighbouring sounds influence one another, can be modelled as part of a post process [104], but

there is no well defined model of coarticulation in the phoneme-to-viseme mapping. Also, the boundaries between the units in the acoustic and visual modalities are assumed to align and in general this is not true. In a standard phoneme-to-viseme mapping there is no accounting for the natural asynchrony of audiovisual speech. More significantly, and more seriously, a phoneme is by definition a group of related **sounds** that are perceived to have the same function. Phonemes serve to represent meaningful contrasts between acoustic speech utterances. Different realisations of the same phoneme can, and often do, appear very different visually.

This thesis introduces a new, dynamic viseme which represents contrastive movements of the speech articulators that are derived by analysing real *visual* speech, rather than by clustering phoneme labels. Dynamic visemes better represent the visual speech signal in that each viseme serves a particular function, and so substituting one dynamic viseme for another changes the visual meaning of the utterance. The dynamic nature of the unit means that coarticulation effects are explicitly modelled, and the boundaries between visemes are not tied to the boundaries of the underlying phones. Indeed, as the units represent the **movements** of the visible articulators, a single dynamic viseme typically extends over several acoustic phones, so the relationship between phoneme sequences and dynamic visemes is complex and many-to-many.

Dynamic visemes are learnt by clustering visual speech gestures in a large corpus of video data. A gesture is defined as a short, intuitive movement of the articulators, and is determined by automatically segmenting the visual speech based on the dynamics of the articulators. The gestures are clustered such that those appearing within a dynamic viseme class portray the same visual function, and represent the visual equivalent of the allophones of a phoneme. An overview of the training process is shown in green in Figure 1.1.

To animate new speech, the phonemes corresponding to clustered gestures are searched to determine a set of candidate dynamic viseme sequences that might produce the desired utterance, and a cost function selects the best sequence to use.

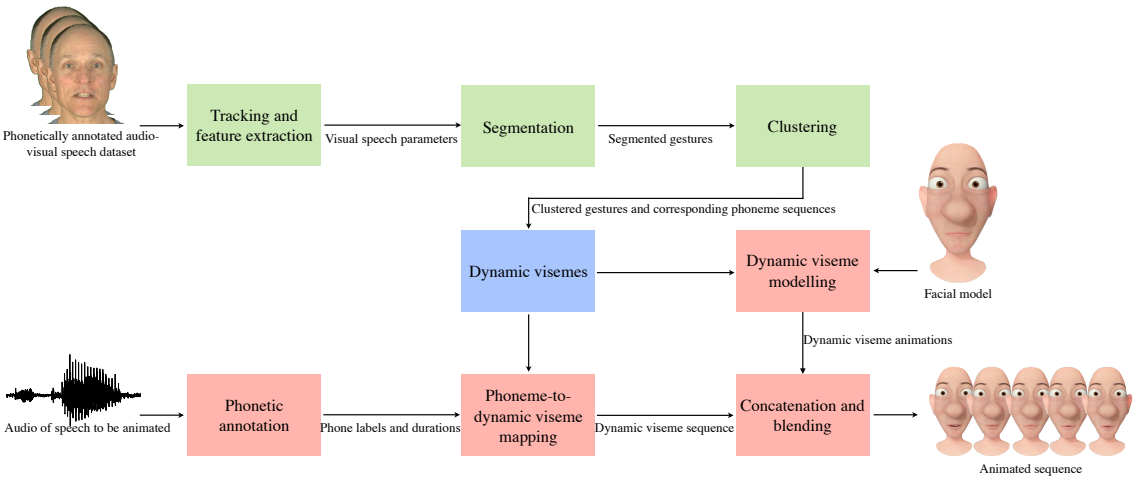


Figure 1.1: An overview of the processes involved in defining dynamic visemes (green), and applying them to speech animation (red).

Since each dynamic viseme cluster represents a particular movement on the lips, speech animation can be generated by modelling one example from each cluster on a facial model, and then retiming and stitching together the viseme clips to form the required sequence. The animation pipeline is shown in red in Figure 1.1.

1.1 Contributions

- The design, capture and annotation of a large audio-visual speech dataset.
- A novel approach at segmentating visual speech based on the dynamics of the visual articulators.
- The definition of a novel, dynamic visual speech unit derived from analysis of real speech.
- A proposed mapping between acoustic and visual speech units.
- Formal objective and subjective evaluation of speech animation.

1.2 The Importance of Visual Speech

The acoustic signal is the primary modality of speech and is usually sufficient for everyday communication, as illustrated by the use of the telephone and radio. For a long time, speech was thought to be a unimodal, auditory process and the visual counterparts were merely considered a bi-product of moving the articulators to produce the acoustic targets. It is now known that the visual cues also provide important information, which the brain uses to decode speech more reliably. Visual speech considers the position and dynamics of the articulators that are visible during speech. These are usually assumed to be the lips, jaw and often the tongue and teeth, but might also include the throat and cheeks, as these also move as a direct result of speech and convey information regarding the content of an utterance.

Visual speech information enhances the intelligibility of speech for hearing impaired people, and it has also been found to improve intelligibility for people with normal hearing in acoustically noisy environments, such as a bar or factory [134, 139]. Where background noise makes speech difficult to hear, a large amount of information is exchanged via the visual modality. Sumby and Pollack [139] were among the first to measure the impact of visual speech on people's ability to understand spoken words in noisy environments. Participants were asked to identify words from a closed set under audio-only and audio-visual conditions with varying amounts of added acoustic noise. They found that the presence of the visual signal was approximately equivalent to a 12dB gain in the signal-to-noise ratio (SNR) and that the benefit from the visual modality increases as the information from the auditory modality decreases. Ross et al. [134] performed a similar experiment, however, in their study a closed set of responses was not offered to the participants, increasing the difficulty of the task. They found that although the gain from the visual modality was inversely correlated with the auditory SNR, there was a window around -12dB SNR where bimodal speech processing was especially beneficial.

There is overwhelming evidence to suggest that speech perception is a bimodal process for normal hearing people when presented with clearly audible speech. The

most convincing argument to support this theory is the illusion known as the McGurk Effect, first described by McGurk and MacDonald in 1976 [110]. They discovered that by dubbing the audio of a person uttering a syllable onto the video of a speaker uttering a different syllable, the clip was perceived as a third, different syllable. The classic example is a video containing audio "ba" and video "ga", which was usually perceived as "da". This demonstrates the significance of visual cues for speech intelligibility as both the acoustic and visual information contribute to the perceived sound. This idea is further upheld by Arnold and Hill [1] who discovered that for clearly audible speech, intelligibility significantly increases with the addition of the visual information.

Much of the benefit gained from the visual signal may be explained by the complementary nature of audio-visual speech, as certain speech sounds are difficult to disambiguate acoustically but are distinct visually. For example, /n/, /m/ and /θ/, /f/ are acoustically very similar as they are produced in the same manner, but are visually distinct as the place of articulation differs. Conversely, /p/ and /b/ are acoustically distinct, but are visually confusable.

It is not well understood how humans integrate audio and visual information for decoding speech. However, recent research in the field of neuroscience has indicated that silent speech-reading activates the auditory cortex [22] and it has been suggested that the patterns in the auditory cortex reflect implied auditory information [67]. The extent of the cortex activations as yet remain uncertain, and further research is necessary. However these findings suggest that there may be some integrative process that combines visible and heard speech, further supporting the belief that speech perception is a bimodal function.

Aside from speech intelligibility, the addition of visual information to audio speech has been shown to have beneficial effects in the performance of other linguistic tasks, such as language identification [138], discriminating sounds in other languages [115] and word segmentation [137]. It has also been shown to increase speech intelligibility of clearly audible speech when the subject matter is complex [1].

From a very young age people are exposed to face-to-face communication in day-to-day life and are very quickly able to build a link between what they see and hear. It has been shown that young babies are aware of a relationship between audio and visual speech information [2], and children benefit from visual information when speech is clear and audible, just as an adult would [1]. People are therefore extremely sensitive to information conveyed via the face. Nuances such as a small puff of the cheeks, tilt of the head or protrusion of the lips all serve as clues that contribute to the perception of speech. Human's sensitivity to this visual information makes realistic speech animation difficult because if these cues are missing, or wrong, then viewers will identify this immediately and potentially find the animation distracting. The work described in this thesis goes some way towards solving the problem of animating natural-looking speech by analysing the position and dynamics of the articulators from real data to learn an inventory of motions that can be concatenated to produce speech animation.

1.3 Real-World Applications

The work described in this thesis enables realistic visual speech animation for any given text. In 1995, Toy Story became the first full-length computer animated feature film. Since then, animation has become a dominant part of filmmaking with fully animated characters appearing in films such as Gollum from Lord of the Rings in 2001 and many characters from Avatar in 2009. For the highest realism, the state-of-the-art techniques for speech animation use facial motion capture, where an actor is filmed with markers positioned at various locations on their face. The markers are tracked and the motion is mapped on to the computer-generated character. This is an expensive and time-consuming process and requires the actor to be re-tracked for any modifications to the script.

Computer-generated cartoons are becoming increasingly popular, particularly aimed at children. These cartoons are typically made as a series of episodes, so

a fast method for generating speech animation is necessary. It has been found that children are able to improve their vocabulary by watching an animated character speaking [105]. Increasing the realism of the lip motion in cartoons could have a positive affect on a child's lexicon and pronunciation. Indeed, the use of animated characters for educational purposes is not limited to cartoons and has recently been introduced to the classroom to aid specific learning tasks during speech therapy [12, 28, 38]. In [28] it was stated that an animated character has vast potential in this field as they are "informative, emotional and personable".

In the video gaming industry, facial animation typically requires less realism and more speed, as it is necessary that the face merely looks plausible as long as the scene can render in real time. However, the growth of the computer gaming industry and the capability of modern computers is prompting the development of visually realistic gaming. As an example, the state-of-the-art 2011 game L.A. Noire contains incredibly life-like character animation, which was generated by tracking an actor performing all of the scenes and then playing back the performance on the character. The main drawback of using this method is that the facial animation is inflexible to user interaction as all of the animations are pre-recorded. With a more flexible animation technique the character could respond dynamically to player input and say phrases not previously spoken by the actor.

As technology and computers become more advanced, the cost and ease of using animation as an advertising tool becomes more feasible even for small companies. In recent years, animated characters have been used to advertise products such as meat produce, washing powder and cars. This is an effective form of advertising as the viewer can relate to the anthropomorphic characters as they possess human-like behaviour.

Computer avatars are the graphical representation of a user of a forum or chat room and are traditionally static images. It would provide a richer, and more engaging user experience if avatars were animated and able to speak the latest message. Communication is much more personal if a person or character can be *seen* speak-

ing rather than only heard. As discussed in Section 1.2, people find audible speech more intelligible when visual information is provided, even when the audio signal is clear [1], so the addition of a talking character or avatar displayed during telephone conversations may be beneficial to some people.

1.4 Outline of Thesis

The following 3 chapters contain a selection of background material in the fields of speech production and animation. Specifically, Chapter 2 provides an overview of facial anatomy and the physiology of the articulatory system to explain the underlying processes involved in visual speech production, Chapter 3 describes phonemes, and the traditional definition of visemes as clusters of phonemes and Chapter 4 reviews common techniques for generating speech animation. The capture, annotation and parameterisation of an audio-visual speech dataset is then described in Chapter 5, and the segmentation and clustering of speech gestures to determine a set of dynamic visemes is detailed in Chapter 6. Chapter 7 describes how to generate facial animation for new speech by mapping phonemes to dynamic visemes, and investigates the speaker dependence of the dynamic viseme units. Finally, Chapter 8 concludes the thesis, and outlines further work that can be performed to refine the units and, thus, improve speech animation.

Chapter 2

Visual Speech Production

What we see when a person speaks is determined by a large number of interacting processes. Acoustic speech is produced by pushing air from the lungs through the vocal apparatus which are coordinated appropriately to generate each speech sound. Some of these articulators are visible, including the lips, teeth, tongue and jaw. There are a number of overlapping muscles of differing shape, size and structure located around the face and down the vocal tract which are tensed to varying degrees to configure the positions of the articulators appropriately for each speech sound. The contractions of the facial muscles are restricted by the underlying rigid, bony structure of the skull to which they are attached, and although they are not directly observable, the muscles control the jaw activity and deform the layer of skin that covers the face. These deformations are complex and are governed by the elasticity of the skin and the degree of stress applied by the muscle.

This chapter goes some way towards explaining why we see what we do when a person speaks. A modest overview of facial anatomy and the physiology of the articulatory system is presented to provide an insight into the complexities of speech production and how different categories of sound are formed. The phenomenon of coarticulation is then introduced together with a review of the models of coarticulation from the literature.

2.1 Facial Anatomy

In this section, the anatomy of the face is described including the structure and biodynamics of the skin, the position and function of the facial muscles and the organisation of the skull bones. As the focus of this work is speech, only the lower section of the face is considered.

2.1.1 Skin

Skin is the largest human organ, covering the entire body and serving many important functions. It acts as a waterproof, insulating barrier, protecting against extreme temperatures, and guards the bones, ligaments, muscles and internal organs from injury, drying out and foreign bodies that could cause infection. Human skin is abundant with nerves, cells, and sweat and sebaceous glands and has a layered structure consisting of the epidermis, the dermis and the hypodermis.

The outermost layer is the epidermis, which is a stiff layer consisting mostly of cells made of keratin, a tough protein that is also found in hair and nails. In the lower layers of the epidermis, cells reproduce rapidly and replace the cells from the superficial layers [10]. As the epidermis contains no blood vessels, these cells die off and are completely replaced every 4-5 weeks. The epidermis ranges in thickness from 0.05mm on the eyelids to 1.5mm on the palms and soles of the feet and also contains melanocytes, which forms melanin and gives skin its colour.

Next is the dermal layer, which is around ten times the size of the epidermis and is the tissue that defines the mechanical properties of the skin in terms of elasticity and strength. The dermal tissue contains 72% collagen and 4% elastin fibres which form a ground, incompressible, gelatinous substance which provides low resistance at low stress and higher resistance at high stress. When the skin is stretched, the collagen fibres uncoil in the direction of strain allowing the skin to deform, so when a large force is exerted from the muscle, the fibres fully uncoil and provide a lesser stretch. When the stress has released, the elastin fibres act like springs and return

the collagen fibres to their original uncoiled state. The behaviour of the skin is therefore complex and non-linear [141].

The dermis contains blood vessels, which regulate body temperature, and a network of nerves that sense pressure, pain and temperature and relays them to the brain. It also accommodates hair follicles, sweat glands and sebaceous glands that produce oil, lubricating the skin and hair. As a person ages, the amount of collagen and elastin in the dermis decreases causing skin to be less elastic. Together with an overall decrease in subcutaneous tissue, this encourages skin to sag and wrinkle.

The hypodermis is typically not classified as skin but as a layer primarily consisting of loose connective tissue and lobules of fat. The fibrous connective tissue serves to fasten the skin to the underlying layer of muscles, and the fat provides thermal insulation and acts as a shock absorber for the bones and as a cushion for the skin.

2.1.2 Muscles

The deformation of the skin is largely controlled by the contraction of an aggregation of facial muscles. Muscles can be described as bundles of fibres working in unison [142], where shorter fibres are more powerful and longer fibres have a larger range of movement. Facial muscles generally arise from a bone at one end (the origin) and insert into the skin at the other (the insertion). Muscles pull the tissue to which they are attached towards the bone from which they emerge and often cause wrinkling of the skin at right angles to the contraction. There are two types of muscle contraction: isometric and isotonic. Isometric contraction, literally meaning *same length*, causes muscles to tense without changing size, whereas isotonic contraction allows the muscle to shorten whilst tensing. When shortening, the other dimensions of the muscle increase to maintain a constant volume. Figure 2.1 illustrates the position of the major muscles of the lower facial area.

The circular muscle surrounding the mouth, the orbicularis oris, has the most complex muscular interaction of all facial muscles as it has no attachment to the

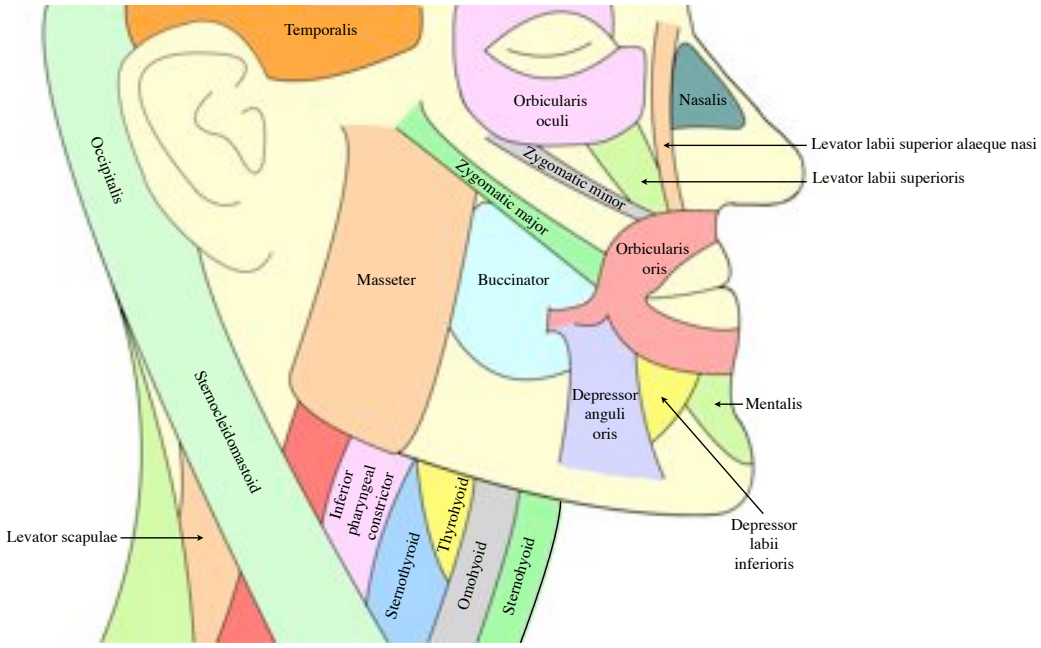


Figure 2.1: The major muscles of the human head.

bone. It consists partly of fibres belonging to the lips and partly of other facial muscles such as the buccinator. This muscle has the primary control over the mouth movements, and is very important for articulated speech as it shapes and controls the size of the mouth opening and is essential for creating the lip configurations necessary for speech. Contraction of the orbicularis oris causes narrowing, rounding and puckering of the lips.

The buccinator controls movement of the anterior portion of the cheek and the lateral wall of the oral cavity. It originates at both the upper and lower jaw and from a fibrous structure extending from the hamulus, a thin, curved bone attached to the sphenoid bone (see Figure 2.2). Contraction of the buccinator pulls back the angle of the mouth and flattens the cheek area enabling the production of sounds such as /i/. Without the buccinator, speech would be difficult and sound slurred.

In close proximity to the buccinator is the depressor anguli oris. This is a triangular muscle originating on the mandible that depresses the angle of the mouth.

A complex combination of muscles are involved with simple opening and closing

of the mouth. The elevation of the upper lip is performed with a combination of the activations of the levator labii superioris, levator labii superior alaeque nasi, zygomatic major and zygomatic minor muscles. Each of these muscles contract at a slightly different angle. The levator labii superior alaeque nasi muscle is also responsible for the dilation of the lateral surface of the external nose and the zygomatic major also pulls the lips laterally. Conversely, the depressor labii inferioris controls the depression of the lower lip, enabling exposure of the lower teeth. The mentalis raises the chin and in doing so, dislodges the lower lip such that it elevates and protrudes.

One of the strongest facial muscles is the masseter. This is a broad, thick, rectangular muscle that originates from the zygomatic bone and inserts into the mandible. The function of the masseter is to elevate and draw the mandible forward, and, in doing so, closes the jaw.

2.1.3 Bones

The skull is the upper-most part of the skeleton and serves as general framework for the head. Excluding the ear bones, the skull is composed of 22 bones [48]; 8 in the cranial area (neurocranium) and 14 in the facial area (viscerocranium). Figure 2.2 illustrates the major bones in the lower portion of the human head.

The cranial area is the section that directly surrounds the brain. It contains many of the larger bones in the head, such as the temporal bones, which form the sides of the facial skeleton and base of the skull, and the frontal bone, which forms the forehead. The sphenoid connects the cranial skeleton to the facial skeleton.

The facial bones help to define features of the face and maintain the rigid structure. The nasal bones form the bridge of the nose and the zygomatic bones form the cheekbones and the lower, lateral eye socket. The maxilla supports the upper teeth and forms the upper jaw, part of the nasal cavity and the anterior section of the hard palate. The posterior portion of the hard palate is formed from the palatal

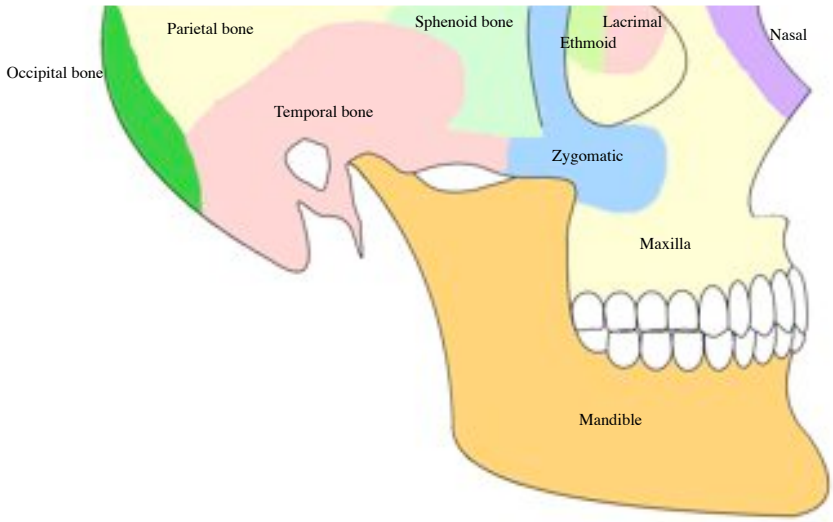


Figure 2.2: The major bones of the head.

bones.

The skull only has one jointed structure, the mandible, which can move vertically and horizontally to some degree and supports the lower teeth. The movement of the mandible enables a chewing motion and is essential for speech production.

2.2 Physiology of the Articulatory System

Speech production is a complex function that involves coordinating the articulators such that the air from the lungs forms a pressure wave that humans perceive as speech. The jaw, lips, tongue, velum, larynx and nasal and oral cavities all play their part in this complicated procedure. For even a simple, one-syllable word, over 70 muscles and the movement of 8–10 body parts between the diaphragm and the lips are required [60]. Figure 2.3 shows some of the articulatory organs used for speech production.

During respiration, the vocal cords are relaxed, allowing air to freely pass down through the trachea and into the lungs and vice versa. To produce *voiced* speech, for example the sounds /v, z, d/, the vocal cords vibrate and interrupt the airflow

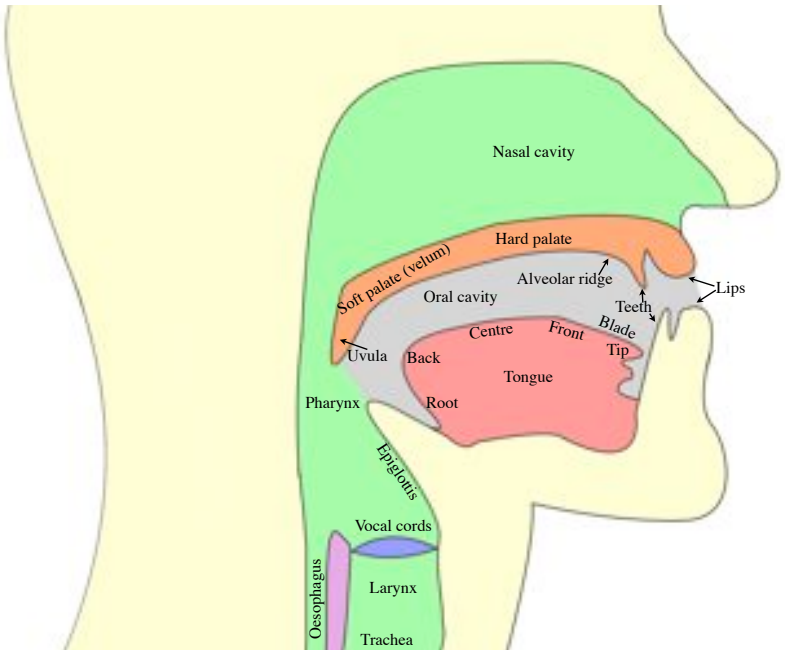


Figure 2.3: A selection of the articulators used for speech production.

through the vocal tract, producing pulses of air. It is the frequency of this vibration that gives pitch to the sound. The vocal cords also have the ability to spread, allowing air to pass through the glottis at high speeds without vibrating the folds, producing noise-like turbulence which is perceived as *voiceless* speech, for example the sounds /f, s, t/.

As air is forced through the vocal cords, a series of constrictions are made by a combination of the articulatory organs to produce different sounds. For example, a bilabial consonant is produced by bringing the upper and lower lips together for the beginning of the sound (for example, the /b/ in "bat") and a labiodental consonant is produced by bringing the lower lip to the upper, front teeth (for example, the /f/ in "fat"). The location of the constriction is referred to as the *place* of articulation. A selection of these are outlined in Table 2.1 [87]. Note that the labio-velar sounds are described as having lip rounding — a pose consisting of tense, protruded lips forming a narrow circular opening.

For some places of articulation, the *manner* in which the sound is produced can

Place of Articulation	Description
Bilabial	Upper and lower lips brought together
Labiodental	Lower lip against upper front teeth
Dental	Tongue tip/ blade against upper front teeth
Alveolar	Tongue tip/ blade against alveolar ridge
Palato-/ Post-alveolar	Tongue blade against back of alveolar ridge
Palatal	Front tongue against hard palate
Velar	Back tongue against soft palate
Glottal	Constriction in glottis
Labio-velar	Back tongue against soft palate and lips brought together and rounded

Table 2.1: A description of the places of articulation for the speech sounds used in this thesis. See Table 3.1 for the list of corresponding phonemes.

vary. For example, the *plosive* (or stop) alveolar sound /t/ (as in "tat") is produced by first completely obstructing the air stream by closing the articulators to build up pressure and then releasing to produce a short burst of sound. Whereas the *fricative* alveolar sound /s/ (as in "sat") is produced by narrowing the articulators, creating a turbulent airflow and a hiss-like sound. Manners of articulation are outlined in Table 2.2.

Articulation of a consonant can always be described in terms of the *voicing*, *place* and *manner*. The consonants used in this work together with an example of their use and a description of how they are produced are shown in Table 3.1.

The articulation of vowels is somewhat different to that of consonants. During the production of vowel sounds, the articulators remain apart, resulting in an unobstructed airflow. This means that vowel sounds are always voiced and none of the articulators constrict the airflow, so we cannot distinguish vowels in terms of voicing, place and manner of articulation. Instead, the cardinal vowel system [76]

Manner of Articulation	Description
Plosive/ Stop	Articulators are completely closed and then released to produce a burst of sound. An oral stop is a closure where both the nasal and oral tracts are blocked off, whereas for a nasal stop only the oral tract is obstructed.
Nasal	Airflow through the mouth is obstructed and redirected through the nose.
Fricative	Air is forced through narrowed articulators, producing a turbulent airflow.
Approximant	Articulators are narrowed, but to a lesser degree than the narrowing necessary to produce fricatives.
Lateral approximant	As above, but in this case the airstream is directed over the sides of the tongue rather than the middle.
Affricate	Affricates begin as plosives and release as fricatives.

Table 2.2: A description of the manners of articulation for the speech sounds used in this thesis. See Table 3.1 for the list of corresponding phonemes.

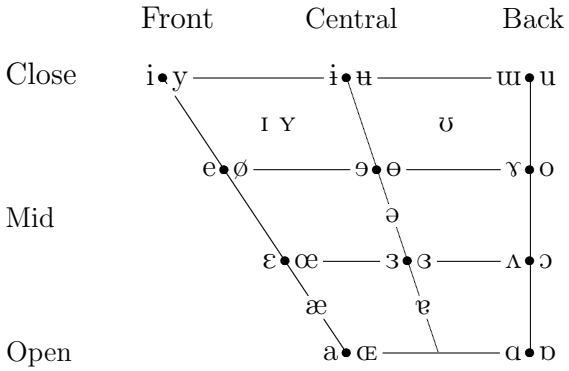


Figure 2.4: The IPA cardinal vowel chart, where the rows describe the vowel height and the columns decribe the vowel backness. See Table 3.2 for examples of the vowels embedded in words.

was introduced to describe vowels in terms of tongue height, backness and lip roundedness. Figure 2.4 illustrates the cardinal vowel diagram where the rows describe the vowel height and the columns describe vowel backness.

In this diagram, vowel height roughly relates to the vertical position of the tongue and was determined by analysis of the relative frequency of the first formant (F1) where the higher the F1 value, the more open the vowel. Close vowels such as /i/ and /y/ are produced with a high tongue and close jaw positions, whereas open vowels such as /a/ and /ɶ/ are produced with a low tongue and an open jaw position. Vowel backness roughly relates to the horizontal position of the tongue relative to the front or back of the mouth. During the articulation of front vowels such as /e/ and /ø/ the tongue is positioned at the front of the mouth, near to the teeth, whereas for back vowels such as /ɯ/ and /u/ the tongue is positioned at the rear of the mouth. Where symbols are shown in pairs, the first represents the unrounded vowel and the second represents the rounded vowel. The vowels used in the work described in this thesis are listed in Table 3.2 with an example word that exhibits the corresponding sound.

Interestingly, when unable to realise an articulatory target due to obstructions (such as a pipe, or a clamp on the jaw), speakers can and do successfully compensate by using different articulators to approximate the sound [95]. Indeed this means that a sound can often be produced with a number of different configurations. An

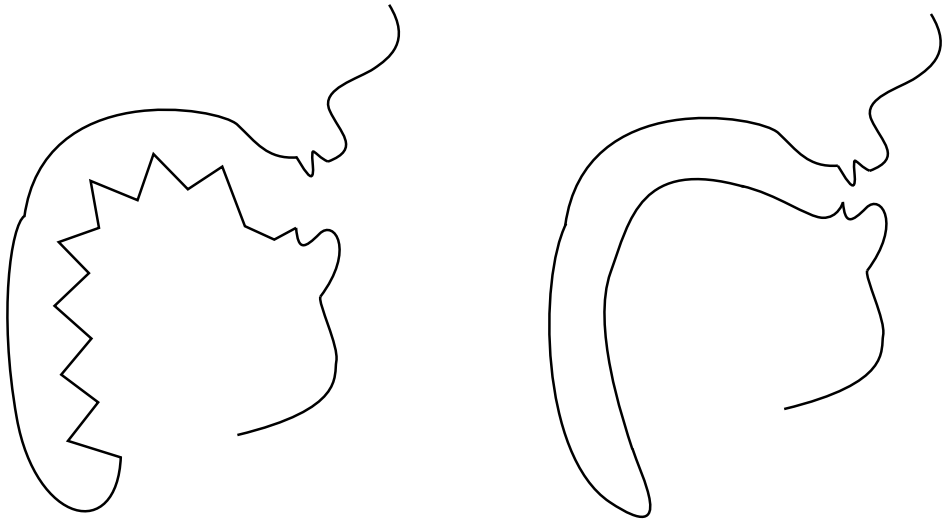


Figure 2.5: The synthetic vocal tract position on the left produces the same first three formants as that on the right, illustrating how the same sounds are produced with a variety of articulator configurations. This is a reproduction of an image taken from page 1 of Ladefoged et al. [86].

amusing example taken from [86] is shown in Figure 2.5. In this example, the synthetic vocal tract shape on the left produces the same first three formants as that on the right.

2.3 Stress

Stress is a linguistic feature that is used to add emphasis to a syllable or word. A stressed syllable is perceptually more salient than an unstressed syllable, as it is typically of higher (or lower) pitch and increased duration, and is produced by pushing more air from the lungs than the surrounding sounds. It was found that /ɔ/ and /e/ always bear stress in open and closed syllables, /i/ and /u/ in closed syllables and /i/ and /u/ can either be stressed and unstressed in open syllables [113] where open and closed syllables are single vowel syllables that respectively end in either a vowel or consonant.

Stress is a suprasegmental feature of speech, as it applies to syllables rather than individual phones. It is an important aspect of the English language (and other

Germanic languages), as varying the stressed syllable in a word can, and often does influence the perceived meaning of an utterance. For example, "to insult" versus "an insult" [87]. Another example is contained within the sentence "I never said she stole my money". If this sentence is read aloud seven times, each time stressing a syllable in a different word, it takes on seven different meanings. It is common for more than one word in a sentence to contain a stressed syllable, and for longer words to have a primary and a secondary stressed syllable.

English is considered to be a stress-timed language, where, during an utterance, the stressed syllables are spaced uniformly in time. However, this is largely disputed as the rhythmic timing of English is thought to depend on many interacting features [132].

2.4 Audio-Visual Asynchrony

It is intuitive that the articulators need to be appropriately positioned prior to the acoustic onset of a phone. Therefore, visual speech typically precedes acoustic speech by tens to a few hundred milliseconds [148] and the articulatory period of visual speech is longer than the acoustic [6]. Bregler and Konig [17] observed that, upon inspection of the mutual information between acoustic and visual features with varying temporal offsets, acoustic features were most correlated with visual features 120 milliseconds in the past.

Audio-visual asynchrony is incredibly complex as the timing of different articulators varies [6]. It is not known whether this is attributable to motor planning or the biomechanics of the articulators.

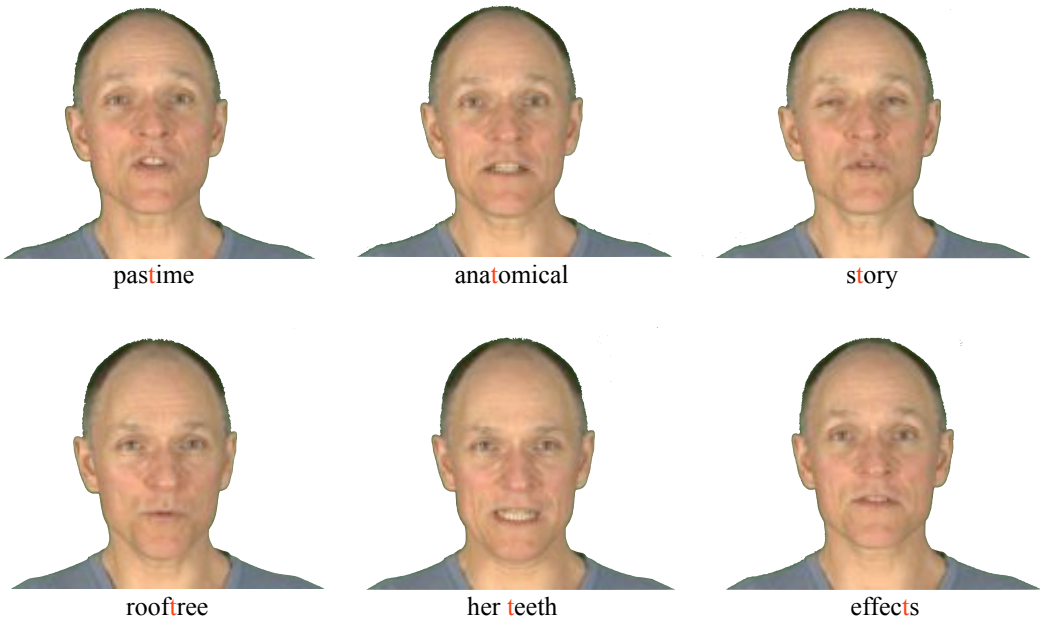
Audio-visual synchronisation plays an important role in sound source location by humans, as sounds are perceived to originate from the stimuli that is synchronisation with the audio. This is particularly apparent in the case of a ventriloquist's dummy or a television screen and is known as the *ventriloquist effect* [65]. This effect has been exploited for computer-based sound source localisation, to determine which

person in a camera shot is speaking [65].

2.5 Coarticulation

Coarticulation is the influence of neighbouring speech sounds on the configuration of the articulators. A particular sound can often be produced with a range of different visible configurations with minimal effects on their auditory characteristics, as illustrated in Figure 2.6 which shows example video frames that were extracted midway through the production of the phonemes /t/ (top) and /k/ (bottom) embedded in sentences. This image shows significant variation in the pose of the articulators as the lips are spread during the /t/ in "teeth" and rounded during the /t/ in "story". This variation occurs because only a subset of the articulators are required to produce each sound. The redundant articulators often remain at the position of a previous sound, or move early towards the next configuration that requires them. Segments of speech are therefore highly influenced by the surrounding context and segment boundaries are visually blurred. As an example, lip rounding is necessary for producing the sound /w/ in the word "twig". However, due to coarticulation, the preceding /t/ also appears rounded. Coarticulation is thought to be caused by a combination of motor planning, the constraints of the human muscle system, linguistic contrast and effort minimisation

To measure the effect of coarticulation in the acoustic modality, it is common practise to calculate the change in first and second formant frequencies (F1 and F2) of the vowels in varying contexts [4, 102, 103, 113]. Vowels typically have over four distinguishable formants, however, it is thought that the first two formants are sufficient to determine the quality of the vowel sounds, as F1 describes the open and close dimension and F2 describes the front and back dimension [87] (see Figure 2.4 for the vowel chart). Using this method, studies have shown that the influence of coarticulation is bidirectional. Carry-over (or backwards) coarticulation is thought to reflect biomechanical and inertial limitations [64, 113] and describes



(a) /t/

(b) /k/

Figure 2.6: A selection of movie frames during the articulation of the phones /t/ and /k/, illustrating the variability of articulator poses due to the coarticulation.

the influence of the preceding speech segments on the articulators. For example, the lips are spread during the /t/ that appears at the end of the word "neat". However, when uttering the word "naught", the lips are rounded during the /t/ sound. Conversely, anticipatory (or forwards) coarticulation describes the influence of the following speech segments on the articulators and is thought to be attributable to pre-programming strategies [64, 113]. As an example, during the words "deem" and "doom" the anticipatory spreading of the "ee" (/i/) and rounding of the "oo" (/u/) significantly affects the shape of the lips during the /d/ sound.

It is generally agreed that coarticulation influences are also asymmetrical in terms of direction. For example, Beddor et al. [4] found that the English language has a larger carryover influence than anticipatory. For American English, Modarresi et al. [113] analysed the bidirectionality of coarticulation within $CV_1.CV_2$ and /tV_1C.V_2t/ utterances by measuring the change in F2. Their findings suggest that for closed syllables carry-over coarticulation is significantly greater than anticipatory, and for open syllables anticipatory coarticulation is significantly greater than carry-over, but not for the vowels /i/ and /e/. This indicates that it is much more likely that the asymmetry of coarticulation is a function of phonetic context rather than a fixed rule for a particular language.

Coarticulation effects have also been found to be context dependent. Fowler and Saltzman [53] asserted that this is clearly the case as, to produce a particular gesture, the articulators that are affected by coarticulation need to be those that do not interfere with the achievement of the gestural goals, or only interfere within a tolerable amount. They will therefore be different in varying contexts.

The number of phonemes in a language's inventory varies from ≈11 in Pirahã, spoken by a tribe in Brazil, and Rotokas, spoken in a small island in Papa New Guinea, to ≈93–111 in Taa, spoken in parts of Africa. Unsurprisingly, given the vast number of language-specific phonemes, coarticulation effects have also been found to be language-specific [4, 103]. For example, Beddor et al. [4] investigated the difference in vowel-to-vowel coarticulation across English and the African lan-

guage Shona by analysing the first three formants of $CV_1CV_2CV_3$ words. They found that in Shona, the anticipatory coarticulation was greater than carry-over, whereas in English, carry-over effects were at least as large as anticipatory. English carry-over effects were measured to be, on average, 2.4 times the length of those in Shona. These findings suggest that coarticulation is not merely a mechanical artefact resulting from acoustic speech production, but it is a learnt process that differs across languages. Manuel [103] theorises that this is due to each of the phonemes having an associated tolerance that defines how far from the ideal target the articulators are allowed to stray. These tolerances are different for each language based on the distinctiveness of the phonemes to one another. According to this theory, a language that has a large number of phonemes is produced with less coarticulation than that with fewer phonemes, as the phonemes are less distinct from one another. Upon analysing the first and second formants of the acoustic speech for a selection of African languages with varying phoneme inventories, this was found to be the case, as vowels from languages with a small inventory were found to be more influenced by anticipatory coarticulation. Results like these indicate that a theory of coarticulation must account for psychological as well as physiological constraints [133].

To measure the effect of coarticulation in the visual modality, which is the influence of coarticulation on the *visible* articulators, video is directly analysed using computer vision or subjective approaches. Computer vision based approaches extract visual information and process this signal [7, 147], whereas subjective methods involve recording participants' responses to a lipreading task [8].

Coarticulation is not merely a function of the directly neighbouring speech units. Instead, by analysing the signal taken from a photocell, Benguerel and Cowan [7] discovered that anticipatory protrusion in French vowels may occur up to six speech units before the vowel is realised. It is thought that each phoneme has an associated visual dominance that controls both the degree of influence it has on the adjacent and near adjacent units, and how far the coarticulation effects spread. A phoneme's dominance and deformability depends on whether fully reaching the articulatory

targets is necessary to produce the required sound. This means that not all visual phones are equally affected by coarticulation as the organs that are deemed necessary for producing a sound may or may not be visually apparent. For example, the consonants /f/ and /v/ are far less deformable than /k/ and /g/. The former are labiodental consonants that are articulated using the upper teeth and lower lip — granting minimal freedom to the shape of the lips — whereas the latter are velar consonants that are articulated at the back of the soft palate, granting more freedom to the contour of the lips.

To determine which of the vowels are most dominant, Owens and Blasek [119] and Benguerel and Pichora-Fuller [8] performed consonant recognition tasks and found lowest lip-reading accuracy when consonants were followed by /u/, suggesting that this is a more visually dominant vowel as lip rounding and protrusion are both essential to produce the sound. This result is accordant to the results of Turkmani et al.'s experiment [147] in which a parameterisation of the visual speech was extracted by tracking the lip boundary and applying PCA to the feature points. By applying linear discriminant analysis to the first two principal components, they found that for VCV words where V = {/i/, /ʌ/, /u/} and C = /p/, the utterances with the /u/ vowel context were most dissimilar to the others with respect to both shape and timing. Benguerel and Pichora-Fuller [8] also found that in VCV contexts /u/ attained a near perfect recognition score whereas /æ/ scored the lowest. Perkell and Matthies [126] measured coarticulation in /iCu/ utterances by recording vertical displacement of a point on the upper lip. They found that many of the subjects began lip-protrusion for /u/ directly after the acoustic offset of /i/.

There are other factors that contribute towards the degree of coarticulation in an utterance. The articulators move extremely rapidly when an utterance is spoken at a normal rate [87]. When speaking at a faster rate a person's lips move less [123], so coarticulation effects increase [85]. This is unsurprising as instantaneous transitions between articulatory targets are impossible, so overlapping and merging of the speech segments is inevitable. Lexical stress also influences the degree of coar-

ticulation. Beddor et al. [4] found that stressed syllables are far less influenced by coarticulation effects and Fowler [52] discovered that a stressed vowel exerts more coarticulation influence on the surrounding vowels. However, these findings conflict with those by Magen [102] who measured a stronger effect emerging from stressed vowels in only one of four speakers. Magen also found significant differences in the degree and direction of coarticulation across speakers uttering words in the form /bV$_1$bəbV$_2$b/.

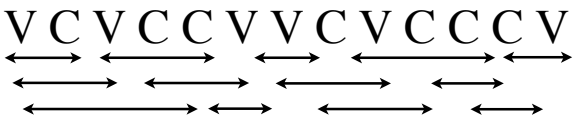
2.5.1 Modelling Coarticulation

If coarticulation is not accounted for, speech animation will appear unnatural and unrealistic. This is because linguistic units, such as phonemes and syllables, are variant in terms of appearance as the position of the articulators during speech is determined by factors such as phonetic context, rate of speech, language, physical constraints and prosody, which all contribute towards the non-symmetrical, bi-directional phenomenon of coarticulation. This makes visual speech a difficult process to model. Models of coarticulation are typically mathematical or rule-based, and account for the direction and duration of neighbourhood influence, and the outcome of the gestural conflict to determine the appropriate composition of the articulators after competing coarticulation influences are imposed.

The simplest coarticulation models assume that the degree of influence spans only one unit either side of a speech segment. For example, Wickelgren [154] claimed that speech was organised in terms of context sensitive allophones rather than phonemes, in which the context encompassed only the directly adjacent phonemes. This effectively converts the phoneme sequence /ð ʌ f r ɛ t f ʊ l ɛ l k/ ("The fretful elk") into the allophone sequence /$_{\#}$ð$_{\text{ð}}$ʌ$_{\text{ʌ}}$f$_{\text{f}}$r$_{\text{r}}$ɛ$_{\text{ɛ}}$t$_{\text{t}}$f$_{\text{f}}$ʊ$_{\text{ʊ}}$l$_{\text{l}}$ɛ$_{\text{ɛ}}$l$_{\text{l}}$k$_{\text{k}}$k$_{\#}$/ [154]. Although simple, this method fails to account for long term coarticulation effects.

Most coarticulation theories tend to be grouped into one of two categories, feature spreading or co-production models as illustrated in Figure 2.7. This section explains these models and Chapter 4 describes a selection of coarticulation models that have

(a) Feature spreading



(b) Co-production

Figure 2.7: Two common coarticulation theories. The direction and length of the arrows represent the influence of coarticulation and C and V represent consonants and vowels respectively. (a) Feature spreading model illustrating the C_nV segment structure proposed by Kozhevnikov and Chistovich [84], where the consonants are assumed to be non-bilabial. In this model, all non-bilabial consonants that directly precede a vowel are influenced by the vowel, regardless of duration and number of units. (b) Co-production model where the gestures corresponding to each phoneme overlap one another. The overlap can be symmetrical or asymmetrical and varies in length according to the dominance of the phoneme.

been implemented to generate speech animation.

2.5.1.1 Feature Spreading Model

Feature spreading (or look-ahead) models, such as Henke's [64], regard speech production to be organised in terms of sequences of discrete, planned, non-overlapping bundles of features that represent the canonical spacial targets of each phoneme. To avoid abrupt changes in the configuration of the articulators, the transitions between adjacent segments are smoothed [51]. If an articulator performs no action for a particular phoneme, it anticipates the value of the next phone in which it necessitates the production of the phone. This model assumes that coarticulatory spread is timeless, as the articulators are assumed to be positioned regardless of the number of phones in the future that are next required.

A similar model by Kozhevnikov and Chistovich [84] asserted that Russian speech is organised in terms of articulatory syllables with a C_nV structure. That is, sequences containing any number of consonants followed by a vowel. As illustrated

in Figure 2.7(a), it is postulated that every non-labial consonant directly preceding a vowel is influenced by the vowel, regardless of temporal duration and number of units. They theorised that motor control is discontinuous at vowel boundaries and so anticipatory coarticulation is bounded by the vowel. This theory was found to not generalise to American English [80] as it fails to consider VV and VC coarticulation, prevalent features of continuous speech. Kent and Moll [81] used images taken from a fluoroscope to analyse the movements of the jaw and tongue for V_1V_2 and V_1CV_2 sequences both within a word and across word boundaries. They found evidence to suggest that the V_1 is influenced by V_2 in all cases. However, as the sequences are embedded in carrier sentences, it is unclear whether other neighbouring segments had any impact on the observed variation.

2.5.1.2 Co-Production Model

Co-production models explain coarticulation as the overlap of co-produced speech gestures in which the magnitude of each speech segment is modelled as a time varying function, and at any one time the position of the articulators is an aggregation of the overlapping functions. In contrast to the feature spreading approaches, Bell-Berti and Harris [5, 6] asserted that the activation of each articulator occurs at a constant time prior to the production of a phoneme if there is no articulatory conflict, where a conflict is the active involvement of an articulator for production of a target sound. Although this model allows for the onset of anticipatory coarticulation to occur midway through a phone or syllable, it is also based on the assumption that the period of anticipation is temporally independent of the speaking rate. Other studies have reported conflicting results [95], stating that the onset of lip-rounding occurs at a time before the rounded vowel proportional to the duration of the preceding consonant.

Löfqvist's theory of speech production [95] is arguably the most well-known. He asserted that at any one time, the position of the articulators represents an aggregation of gestures associated with the production of different speech sounds.

The influence of each gesture over time is controlled by a dominance function where, initially, the influence of the segment on the articulators is zero, then it gradually becomes more dominant, and eventually decreases back to zero influence. These dominance functions can be asynchronous for different parts of the vocal tract, and some gestures may not affect specific articulators at all. The author postulated that during continuous speech, the speech gestures overlap to different degrees, a phenomenon that is likely to be due to speaking rate.

Jackson and Singampalli [72] examined EMA measurements to quantitively determine the role of the articulators during speech production. For each phone, the articulators were classified as either critical, dependent or redundant. *Critical* articulators are those that play a crucial role in the production of a phone. A *dependent* articulator is one that moves as a consequence of a critical articulator's motion due to physical constraints, and a *redundant* articulator is free to move without influencing the production of a phone and is more prone to coarticulation effects. At the mid-point of each phone the x, y coordinates of seven oral articulators were measured and the distributions were modelled as a Gaussian probability density function (PDF). For each phone, critical articulators were identified based on the Kullback-Leibler distance between the phone's PDF and the overall mean PDF. For example, the y divergence of the upper lip was high for the production of the bilabial phone /b/ indicating that the upper lip position is critical for this phone, and low for the velar /g/. However, the opposite was measured for the tongue dorsum. Dependent articulators were identified by analysing the inter-articulator correlation. For a complete model of articulatory coarticulation, further research would be necessary to determine the role of the articulators over entire phone segments.

Co-production models are typically more complex than feature spreading models since the role of each articulator might be modelled separately for each phoneme. However, as it is possible to train the models using analysis of real speech, a co-production model might more accurately represent the complexities of coarticulation.

2.6 Discussion

Speech production is a complex process involving the combined effort of a large number of muscles from the face down to the chest. As the diaphragm forces air from the lungs through the vocal folds, the vocal tract is configured to produce vowels and constrictions are formed to produce consonants.

The skull forms the rigid foundation of the face, over which an array of overlapping muscles are positioned. The muscles are attached to the skull and insert into the elastic layer of skin that encloses the face. When contracted, the muscles non-linearly deform the skin to generate different facial poses.

After considering the physical design and control of the facial model, the most important consideration for realistic speech animation is the effect of coarticulation, the influence of neighbouring speech segments on the position of the articulators. This influence has been found have asymmetric, bi-directionality, and be speaker, language and context dependent. To date, several coarticulation models have been proposed, but there is yet to be a definitive model that is widely regarded as truth.

Chapter 3

Phonemes and Visemes

Phonemes represent the set of perceptually distinct speech sounds of a language. They are well established and have been used successfully as the basis for acoustic speech recognition [158] and synthesis [69]. For visual speech recognition and animation, a visual analogue to the acoustic phonemes is assumed, whereby each phoneme is associated with a particular configuration of the visible articulators. As a relatively small proportion of the articulatory information is visible during speech, different phonemes can appear similar to one another and a one-to-one mapping between phonemes and facial poses can result in redundancy. The phonemes are therefore typically clustered into visually contrastive groups such that each cluster contains sounds that are visually indistinguishable from one another. Each *cluster* of phonemes is then represented with a configuration of the visible articulators. These many-to-one mappings from phonemes to poses are referred to as visemes, and are assumed to form the building block of visual speech.

As an example, the phonemes /v/ and /f/ have the same place and manner of articulation (labiodental fricative), but the former is voiced and the latter is voiceless (see Section 2.2 and Table 3.1). Voicing is produced by tensing of the vocal cords, causing vibration through the air flow — a process that is not visible. For this reason /v/ and /f/ tend to appear visually similar. The same is true for the phonemes /b/ and /p/. The standard approach is to assume a many-to-one relationship between

the acoustic units and the visual units, and assign /v/ and /f/ to a viseme class and /b/ and /p/ to another.

In this chapter, the concept of phonemes and the phonetic labelling system that is used throughout the remainder of this thesis are introduced. The traditional definition of a viseme is then described and the various methods that have been adopted for obtaining the viseme clusters are reviewed. This chapter ends with a discussion detailing the limitations of the conventional many-to-one relationship between phonemes and visemes.

3.1 Phonemes and Phonetics

The basic unit of acoustic speech is the phoneme. A phoneme is an abstract, linguistic unit that represents a *collection* of speech sounds that are perceived as equivalent. The acoustic realisations of the speech sounds are referred to as phones and the set of phones that form a phoneme class are called allophones. Allophones can be varied, but they exhibit equivalent meaning. For example, the /l/ is often unvoiced during the word "play", whereas it is voiced in the word "lay", but both are perceived as /l/ [87]. A defining quality of phonemes is that replacing a particular phone with another that has a different phoneme label changes the meaning of an utterance.

Phonemes provide an unambiguous representation of the speech sounds of a language and are typically placed within a pair of forward slashes (//). There are several phonetic notation systems, including the International Phonetic Alphabet (IPA), SAMPA, and ARPAbet. As IPA is an internationally accepted representation of the speech sounds, this is the notation used where possible in this thesis[1]. IPA was designed such that each distinctive speech sound is represented with a fixed character regardless of language or context [71].

The number of phonemes of the English language ranges from 35 to 47 depending on dialect. A set of 40 phonemes are used in this work, consisting of 24 consonants

[1]In certain cases ARPAbet is used where plain text notation is required.

Place	Manner	Voicing	ARP	IPA	Example
Bilabial	Plosive	u	P	p	apple
		v	B	b	cable
	Nasal	v	M	m	summit
Labiodental	Fricative	u	F	f	fairy
		v	V	v	lava
Dental	Fricative	u	TH	θ	thought
		v	DH	ð	other
Alveolar	Plosive	u	T	t	utter
		v	D	d	udder
	Nasal	v	N	n	canal
	Approximant	v	R	r	hurry
	Fricative	u	S	s	essay
		v	Z	z	zebra
	Lateral approximant	v	L	l	lazy
Palato-/ Post-alveolar	Fricative	u	SH	ʃ	mash
		v	ZH	ʒ	seizure
	Affricate	u	CH	tʃ	such
		v	JH	dʒ	jet
Palatal	Approximant	v	Y	j	yes
Velar	Plosive	u	K	k	track
		v	G	g	bag
	Nasal	v	NG	ŋ	bang
Glottal	Fricative	u	HH	h	hang
Labio-velar	Approximant	v	W	w	worry

Table 3.1: Place, manner and voicing of the English consonants with corresponding ARPAbet (ARP) and IPA symbols and example words that contain the sound. In the voicing column, *u* and *v* denote voiced and unvoiced phonemes respectively.

and 16 vowels. The consonants are listed in Table 3.1 along with their corresponding ARPAbet and IPA symbols.

There are two types of vowel, monophthongs and diphthongs. A monophthong is a single vowel sound where the position of the tongue is somewhat static, whereas a diphthong involves the tongue gliding from one vowel sound to another. A diphthong, such as "show" or "play", is produced with one continuous motion and occurs within a single syllable. The 11 monophthong vowels and the 5 diphthongs used in this work are listed in Tables 3.2 and 3.3 respectively.

Phonemes provide a convenient way of transcribing speech. However, it is also

Closeness	Backness	Roundness	ARP	IPA	Example
Close	Front	Unrounded	iy	i	f<u>ee</u>t
	Back	Rounded	uw	u	b<u>oo</u>t
Near-close	Near-front	Unrounded	ih	ɪ	h<u>i</u>t
	Near-back	Rounded	uh	ʊ	p<u>u</u>t
Open-mid	Front	Unrounded	eh	ɛ	b<u>e</u>tter
	Central	Unrounded	er	ɜ	h<u>ear</u>d
	Back	Unrounded	ah	ʌ	c<u>u</u>p
		Rounded	ao	ɔ	p<u>oor</u>
Near-open	Front	Unrounded	ae	æ	b<u>a</u>t
	Back	Unrounded	aa	ɑ	<u>ar</u>m
		Rounded	oh	ɒ	d<u>o</u>ll

Table 3.2: Closeness, backness and roundness of the monophthong vowels in the English dialect with corresponding ARPAbet (ARP) and IPA symbols and example words that contain the sound.

ARP	IPA	Example
ey	eɪ	d<u>ay</u>
ay	aɪ	b<u>i</u>te
oy	ɔɪ	b<u>oy</u>
ow	oʊ	b<u>oa</u>t
aw	aʊ	h<u>ow</u>

Table 3.3: Diphthongs of the English language with corresponding ARPAbet (ARP) and IPA symbols and example words that contain the sound.

possible that they play a more important role in motor planning and the temporal organisation of speech. This theory stems from the phenomenon known as a *spoonerism*, where two phones, or clusters of phones are unintentionally exchanged within a sequence of words. A famous example was spoken by Professor William Spooner, after whom the phenomenon was named, who reportedly once said "You've hissed all my mystery lectures" rather than "You've missed all my history lectures" [56]. These slips of the tongue are important for speech production theories, as the majority of the speech errors are associated with units that are the size a phoneme segment [56], suggesting that the phoneme has a reality as an action unit [62].

An alternative theory is that the syllable is the organisational unit used in motor planning as Fromkin observed that the exchanged segments of a spoonerism retain their position within a syllable [56]. That is, segments that appear at the beginning, middle and end of a syllable are exchanged with other segments that appear at the beginning, middle and end of a syllable respectively. A syllable is a suprasegmental unit, spanning one or more phonemes. Syllabic boundaries are ambiguous in many cases as phoneticians disagree on the correct segmentation for certain word. For example, the word "puppy" could be segmented into the two syllables "pu-ppy" or "pupp-y". For the English language it is generally agreed that a syllable contains a vowel at its nucleus and optionally, one or more consonants at the boundaries.

3.2 Visemes as Phoneme Clusters

The term *viseme* was originally coined by Fisher in 1968 as an amalgamation of the words "visual" and "phoneme" [49]. Visemes are typically defined as the clusters of visually contrastive phonemes, such that the phonemes that appear within a viseme group are considered visually indistinguishable from one another. This mapping is either assembled by eye [47] or based on the place of articulation and the extent of lip rounding [90, 116]. More commonly, the mapping is defined by clustering the phonemes based on the confusions from a subjective phone recognition experiment

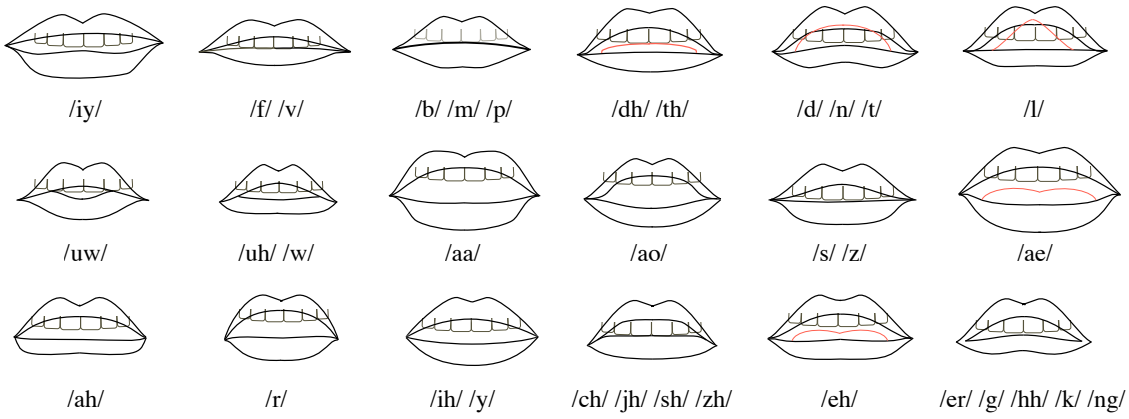


Figure 3.1: The 18 visemes as determined by Parke and Waters [121].

with human viewers or an objective, data-driven visual speech classifier. Tables 3.4 and 3.5 show a subset of phoneme-to-viseme mappings proposed in the numerous studies. Some of the groups formed correspond roughly to clustering on the place of articulation, such as /p, b, m/ and /f, v/, but this is not always the case as evident with larger groups such as /t, d, s, z, n, k, g, j/ [92] and /y, l, n, k, g, h/ [75]. The 18 viseme groups as determined by Parke and Waters [121] along with the associated lip poses are shown in Figure 3.1.

3.2.1 Subjectively Defined Phoneme-to-Viseme Mappings

Early methods for finding the mapping from phonemes to visemes were typically subjective, based on analysing confusions within 'stimulus-response' matrices [3, 11, 49, 91, 92, 94, 152]. Clusters were accepted as visemes when the within-cluster response made up a high percentage of the responses. As an example, Lesner et al. [92] implemented a hierarchical clustering algorithm that incrementally grouped the consonants that were consistently confused with other consonants by participants during a lipreading study with nonsense words in the form /ɑCɑ/. If a consonant was never confused with another, or was not consistently confused with another, it constituted a viseme class of its own. The algorithm converged when the within-cluster response made up 75% of the responses, generating seven consonant viseme

Author	Condition	Viseme groups
Fisher (1968) [49]	Five word phrases	
	Initial	/p, b, (m, d)/ /f, v/ /k, g/ /ʍ, w, (r)/ /ʃ, t, (n, l, s, z, dʒ, j, h)/
	Final	/p, b/ /(f, v)/ /(k, g, ŋ, m)/ /ʃ, ʒ, dʒ, (tʃ)/ /t, d, n, θ, ð, z, s, r, l/
		(Items in parentheses are directional confusions)
Franks and Kimble (1972) [55]	/C*ʌ/	/dw, gw, kw, sw, tw, skw, w, wh, r, dr, fr, gr, kr, ʃr, tr, skr, spr, str/
		/bl, br, b, m, pl, pr, ʃm, sm, sp, spl/ /gl, kl, sl, l, d, sk, sn, st, s, θr, θ, t/ /ʃn, ʃ, tʃ, dʒ/ /fl, fr, f/
Binnie et al. (1976) [11]	/Cɑ/	/p, b,m/ /f, v/ /θ, ð/ /w/ /r/ /t, d, s, z/ /l, n/ /k, g/
Walden et al. (1977) [152]	/Cɑ/	
	Pre-training	/b, p, m/ /f, v/ /θ, ð/ /s, z, ʒ, ʃ/
	Post-training	/b, p, m/ /f, v/ /θ, ð/ /s, z/ /ʒ, ʃ/ /t, d, n, k, g, j/ /w/ /r/ /l/
Lesner and Kricos (1981) [91]	/hVg/	
	Speaker 1	/aɪ/ /ɔɪ/ /i, ɪ/ /e, ɛ, æ/
	Speaker 2	/aʊ/ /o/ /i/ /e, ɛ, æ/ /a, ɔ/
	Speaker 3	/aɪ/ /o/ /i, ɪ/ /aʊ/ /ɔɪ/
	Speaker 4	/i, ɪ, ʌ/
Owens and Blazek (1985) [119]	/æCæ/	/p, b, m/ /f, v/ /θ, ð/ /w, r/ /tʃ, dʒ, ʃ, ʒ/ /k, g, n, l/
	/ʌCʌ/	/p, b, m/ /f, v/ /θ, ð/ /w, r/ /tʃ, dʒ, ʃ, ʒ/ /t, d, s, z/
	/iCi/	/p, b, m/ /f, v/ /w, r/ /tʃ, dʒ, ʃ, ʒ/ /t, d, s, z/
	/uCu/	/p, b, m/ /f, v/
Lesner et al. (1987) [92]	/ɑCɑ/	/p, b, m/ /f, v/ /ʃ, ʒ, dʒ, tʃ/ /w, r/ /l/ /t, d, s, z, n, k, g, j/
Goldschen (1994) [58]	Sentences	
	Consonants	/p, b/ /b_{cl}, m, p_{cl}/ /tʃ/ /d, d_{cl}, g, g_{cl}, k, k_{cl}, l, n, t, t_{cl}/ /ð/ /f, v/ /h/ /dʒ/ /ŋ/ /r/ /s, ʃ, z/ /θ/ /w/ /j/ /ʒ/
	Vowels	/a/ /æ, ɛ/ /ʌ/ /ɔ/ /aʊ/ /ə, ɪ, i/ /aɪ/ /ɜ/ /e/ /o/ /ɔɪ/ //ʊ/ /u/
		(*cl* denotes closure)
Parke and Waters (1996) [121]	Unspecified	/p, b, m/ /f, v/ /t, d, n/ /l/ /s, z/ /θ, ð/ /r/ /tʃ, dʒ, ʃ, ʒ/
		/k, g, h, ŋ, ɜ/ /ɪ, j/ /ʊ, w/ /i/ /u/ /a/ /ɔ/ /æ/ /ʌ/ /ɛ/

Table 3.4: A selection of phoneme to viseme mappings from literature.

Author	Condition	Viseme groups
Auer and Bernstein (1997) [3]	/hVg/	/u, ʊ, ɜ/ /o, aʊ/ /ɪ, i, e, ɛ, æ/ /ɔɪ/ /ɔ, aɪ, ə, a, ʌ, j/
	/Cɑ/	/b, p, m/ /f, v/ /l, n, k, ŋ, g, h/ /d, t, s, z/ /w, r/ /ð, θ/ /ʃ, tʃ, ʒ, dʒ/
Ezzat and Poggio (2000) [47]	Isolated words	
	Consonants	/p, b, m/ /f, v/ /t, d, s, z, θ, ð/ /w, r/ /tʃ, dʒ/ /k, g, n, l, ŋ, h, j/
	Vowels	/ɪ, i/ /ɛ, æ/ /ɑ, ɒ/ /ʌ/ /ɜ/ /ɔ/ /ʊ, u/ /aʊ/ /oʊ/
Neti et al. (2000) [116]	Continuous speech	
	Consonants	/s, z/ /t, d, n/ /ʃ, ʒ, tʃ, dʒ/ /p, b, m/ /d, t/ /f, v/ /ŋ, k, g, w/
	Vowels	/ɔ, ʌ, a, ɜ, ɔɪ, aʊ, h/ /u, ʊ, oʊ/ /æ, ɛ, eɪ, aɪ/ /ɪ, i/
Jiang et al. (2002) [75]	/Cæ/	/p, b, m/ /f, v, r/ /θ, ð/ /w/ /d, t, s, z, tʃ, dʒ, ʒ/ /j, l, n, k, g, h/
	/Ci/	/p, b, m/ /f, v/ /θ, ð/ /w, r/ /s, z, tʃ, dʒ, ʒ/ /j, l, n, k, g, h, t, d/
	/Cu/	/p, b, m/ /f, v, r/ /θ, ð/ /w/ /d, t, s, z, tʃ, dʒ, ʒ/ /j, l, n, k, g, h/
	Combined	/p, b, m/ /f, v, r/ /θ, ð/ /w/ /d, t, s, z, tʃ, dʒ, ʒ/ /j, l, n, k, g, h/
Lee and Yook (2002) [90]	Unspecified	/p, b, m/ /f, v/ /t, d, s, z, θ, ð/ /w, r/, /tʃ, dʒ, ʃ, ʒ/ /ɛ, eɪ, æ, aʊ/
		/k, g, n, l, h, j/ /i, ɪ/ /a/ /ʌ, ə, aɪ/ /ɜ, ɔ, ɔɪ, o/ /ʊ, u/
Hazen et al. (2004) [63]	Sentences	
	Consonants	/l/ /b, p/ /b_{cl}, p_{cl}, m/ /s, z, t_{cl} d_{cl}, n/ /tʃ, dʒ, ʃ, ʒ/ /t, d, t, d, g, k/ /f, v/ /g_{cl}, k_{cl}, ŋ/
	Vowels	/ɪ, i/ /ʌ, a/ /æ, ɛ, aɪ, eɪ, h/ /aʊ, ʊ, U, Oʊ, ɔ, w, ɔɪ/
		(*cl* denotes closure)
Lidestam and Beskow (2006) [94]	/ɑCɑ/	/p, b, m/ /f, v/ /d, k, n, ŋ, r, j, g, ʈ, l/
Melenchón et al. (2007) [111]	Spanish sentences	
	Speaker 1	/m, p/ /θ/ /f/ /t/ /n, r/ /s, l, k/
	Speaker 2	/m, p/ /θ/ /f/ /t, s/ /n, l, k/ /r/
	Speaker 3	/m, p/ /θ/ /f/ /k/ /n, r, k/ /t, s/
Zhao and Tang (2008) [160]	Chinese sentences	
	Consonants	/p, b, m, f/ /d, t, n, l, g, k, h/ /j/ /s, z/ /tʃ, ʃ, ʒ, r/
	Vowels	/a, aɪ, aʊ, an, aŋ/ /ɔ, oʊ, eŋ, ʌŋ/ /ɛ, eɪ, ɛn/ /ɪ, ɪm, ɪŋ/ /u/

Table 3.5: Further phoneme to viseme mappings from literature.

classes. In similar experiments, Lidestam and Beskow [94] determined five viseme classes, and Binnie et al. [11] clustered the confused phonemes while the within-cluster response made up 70% of the responses, generating nine viseme classes. The results of these experiments are detailed in Tables 3.4 and 3.5.

Auer and Bernstein [3] applied a similar hierarchical clustering algorithm to the confusions between consonants from /Cɑ/ contexts and vowels from /hVg/ syllables by using lip-reading responses to estimate the visual similarity of the phonemes. In this study, the algorithm converged when a large increase was measured in the average between-cluster distance, as this was assumed to be where two relatively dissimilar clusters were to be merged. Using this approach they identified twelve viseme classes, seven consisting of consonants, and five of vowels. Auer and Bernstein's study was one of a small number that clustered vowels including [58, 90, 91]. The majority of previously determined phoneme-to-viseme mappings assign each vowel its own viseme class as they are considered difficult to group.

Fisher [49] defined a phoneme-to-viseme mapping by asking participants to lip-read the initial and final consonants of a sentence using a forced-error approach where the correct answers were omitted from a closed-set of possible responses. The results, which are shown in Table 3.4, suggest that the viseme groupings for initial and final consonants differ — a factor that is likely to be attributable to coarticulation. He also observed that initial consonants contained directional confusions. For example, /m/ was significantly confused with /b/ but /b/ was not significantly confused with /m/.

Franks and Kimble [55] also observed directional confusions in their consonant cluster recognition task. This experiment involved analysing confused consonant sequences from stimuli of the form /C*ʌ/ where C* represents a sequence of one or more consonants. They found that /spr/ was often confused with /sw/, /sw/ was confused with /sm/ and /tr/ was confused with /fr/, but the confusions were never reciprocated. They also found that sequences of consonants were confused with single consonants 46% of the time. This is an interesting finding which supports the

theories of coarticulation in that it indicates that the articulators do not reach every target during speech and that certain visual phonemes are deformable to the extent that they are not perceived as distinct gestures. If this study were to be extended such that it considered sequences of consonants in multiple contexts during natural speech, it is likely that the confusions would become far more complex.

A disadvantage of using perceptual studies to determine phoneme-to-viseme mappings is that the responses are dependent on the viewer's ability to lip-read and are therefore likely to vary from person to person. Another drawback is that it is necessary for stimuli to be simple. Most studies involved stimuli consisting of nonsense words where phonemes were presented in a single context [3, 91, 92, 114] or a small number of varying contexts [75, 119]. Montgomery and Jackson [114] stated that the context /hVg/ was chosen as it produced minimal coarticulation effects. However, in natural speech, contexts are varied and gestures are coarticulated so the resulting viseme groups may not sufficiently model real speech.

3.2.2 Objectively Defined Mappings

More recently, computer vision and machine learning algorithms have been used to cluster phonemes into viseme classes to overcome viewer bias, and the limitations of the subjective methods in terms of stimuli and phonetic context. These data-driven approaches typically use some form of unsupervised clustering on visual features, where phonemes that are clustered together frequently are said to form a viseme.

Goldschen [58] described a data-driven method for clustering phonemes into viseme classes. A set of static and dynamic lip features were extracted from video of a speaker uttering full sentences. The video was then manually segmented into phonemes and the segmented features were clustered using a hierarchical clustering algorithm with a HMM similarity measure. The resulting visemes appeared to be fairly consistent with results from perceptual experiments such as [49] (see Table 3.4). However, Goldschen introduced the idea of grouping the lip closure and opening of the consonants /b/, /p/ and /m/ separately, forming the groups /b_{cl}, m,

p_{cl}/ and /b, p, r/ where *cl* indicates closure. Hazen et al. [63] performed a similar HMM based clustering using features describing the appearance of the lips. They extended Goldschen's approach by clustering the closure and release independently for all stop consonants.

For Spanish speech, Melenchón et al. [111] extracted features describing the appearance of the mouth during a set of phonetically balanced sentences which were phonetically labelled and segmented. The visual features were clustered into six viseme groups based on the Bhattacharyya distance.

Brooke and Templeton [20] clustered vowels based on the lip height, width and area between the lips independently such that 75% of the examples of each vowel appears within a cluster. They do not specify the configuration of the resulting groups, but across speakers the number of visemes ranges from one to eight.

De Martino et al. [34] introduced context-dependent visemes for speech animation by analysing nonsense words of the structure CV_1CV_2. Context-independent visemes groups were first initialised manually by clustering phonemes based on the place of articulation. During the production of each phone, the frame corresponding to the first stationary point was assumed to represent the articulatory target. Each viseme was then segmented into context-dependent sub-visemes by clustering the x, y, z coordinates of four markers positioned around the lips and jaw for each of the stationary frames. Using this approach, all context-independent consonant visemes were segmented into a minimum of two context-dependent visemes, and all context-independent vowel visemes remained unsegmented other than /i/ which was partitioned into two groups. These results indicate that a more suitable relationship between phonemes and visemes is many-to-many as the phonemes that appear within a viseme class are variable and dependent on the context. If this approach were extended by clustering a richer set of features on continuous speech, a much more complex sub-clustering can be anticipated.

In 2008, Zhao and Tang [160] proposed a method of speech animation using a set of *dynamic* visemes that correspond to the Chinese syllables. First, a standard

clustering approach is used to group the phonemes into viseme classes based on features describing the geometry of the lips at the mid-point of the phone. During this clustering, consonants and vowels are clustered separately, as shown in Table 3.5[2]. Chinese syllables are then transcribed into their corresponding viseme labels. For example, if /b/ and /p/ belong to viseme V_1 and /æ/ and /ɔ/ belong to V_2, /bæ/, /bɔ/, /pæ/ and /pɔ/ would all be transcribed at V_1V_2. In total, the authors found 40 unique dynamic units which they referred to as dynamic visemes. Worryingly, using this method, /f/ appears in the /b, p, m/ viseme suggesting that the features used for clustering do not contain sufficient information to discriminate between bilabial and labiodental gestures. For speech animation this method captures the dynamics of visible articulators within a syllable, but longer term coarticulation effects are not accounted for.

It has been shown that phoneme-to-viseme mappings obtained using objective methods tend to be less reliable than those defined using more traditional subjective methods both for computer facial animation of speech [108] and for visual speech recognition [25]. This is likely because the features that correspond to a particular phoneme will be dispersed across several clusters, and these clusters will not always be composed of the same set of phonemes. Thus the mapping of phonemes-to-visemes is noisy (see Figure 2.6) and a simple many-to-one mapping is not sufficient to model the complex relationship between the visual gestures and the underlying sounds.

3.2.3 Limitations of Visemes for Modelling Visual Speech

As is clear from Tables 3.4 and 3.5, there has yet to be definitive agreement regarding both the number of viseme classes that are required to represent visual speech, and how the set of phonemes map to visemes. It is also apparent that only a small number of studies consider the full range of phonemes in the English inventory, and vowels are often omitted. Thus, to date, the definition of a viseme is informal and

[2]Note that only those phonemes with a corresponding English sound is included in the table.

as a unit of speech for computer facial animation it has been poorly defined.

There are several possibilities as to why the results differ so vastly across the literature, including the nature of the stimuli, quality of the recording (frames per second, pixel density, sharpness etc.), distance to the camera, illumination and visual intelligibility of the speaker for instance. For subjectively determined mappings, the lip-reading ability of the participants biases the mapping, and for objective studies, the features and clustering methods are likely to cause variation.

The configuration of a person's face in terms of size and shape varies considerably across speakers. It is therefore natural that there will also be differences in the way that people speak. This variation contributes towards a person's lip-read-ability and, consequently, the number and composition of viseme classes [20, 91, 111]. This was confirmed by Lesner and Kricos [91] who presented a vowel recognition experiment in which subjects were asked to lip-read /hVg/ nonsense words spoken by four different speakers. They discovered that speakers who were easier to lip-read generally produced a larger number of viseme classes. A similar experiment was performed by Jiang et al. [75] in which participants were asked to lip-read CV syllables uttered by four speakers. They found that viseme classification was both speaker dependent and context dependent with the number of viseme classes varying from four to six. The same variability was measured using objective methods. Brooke and Templeton [20] clustered vowels on a set of geometric parameters extracted from the outline of the lips and teeth for three speakers. They clustered each of the speakers independently and found significant variation in the number and composition of the clusters across speakers. When the phonemes were forced to cluster into six groups, Melenchón et al. [111] found that the frontal consonant visemes, such as /p, m/, /f/ and /θ/, were consistent across speakers. However, they measured weak agreement in the viseme clusters across speakers for non-frontal consonants.

The lip-reading ability of the people used in subjective studies is likely to cause variation in viseme groupings across participants. Furthermore, Walden et al. [152] found that, over a relatively short period of time, a person's lip-reading ability can be

improved with exposure to the speaker. This means that visemes are not consistent even for one person, as they change as fewer visual confusions are made.

Above all else, the variable that is likely to be the most dominant cause of variation across the phoneme-to-viseme mappings is coarticulation. As is clear from Figure 2.6, the same phoneme appears remarkably distinctive in different phonetic contexts. The standard many-to-one relationship between phonemes and visemes is therefore naive as it forces phonemes into clusters of which are assumed to be visually invariant. It is likely to be for this reason that Mattyeyses et al. [108] measured a smaller synthesis error when phoneme units were used rather than visemes for a sample-based visual speech synthesiser, despite having a fewer number of candidate examples available, and that Ramage [130] found no way of clustering confused phonemes of an automatic visual speech recogniser such that 70% of the confusions occur within the viseme classes.

In light of this, it becomes apparent that a many-to-many relationship between phonemes and visemes is necessary to account for phonemic context [73]. A small number of studies have addressed this many-to-many relationship, by clustering context-dependent phonemes [34, 75, 109, 119] or clustering phonemes dependent on their position within a word [49]. However, to date, no many-to-many phoneme-to-viseme mapping exists that accounts for long term coarticulation. A possible reason for this is because it is intractable to cluster phonemes in all contexts as it is necessary to consider combinations of up to six units either side of a phoneme to capture all coarticulation effects.

Although the simplicity of static visemes is attractive for modelling speech, visual speech units are inherently dynamic [6, 21]. The kinematics of the visible articulators are important for speech perception as they help people to distinguish between speech sounds. The consonsants /b/ and /m/ form a viseme cluster in the majority of mappings in the literature as they share a bilabial place of articulation and therefore require lip closure for production. However, the dynamics of these phonemes are very different as /b/ has a plosive manner of articulation and /m/ has a nasal

manner. It has been shown in [136] that lip-readers are able to discriminate between the phonemes /p/, /b/ and /m/ at a better than chance level even when they are spoken in the same context, disproving the traditional definition of a viseme.

There is other information embedded in dynamic speech. For example, Engström [43] found that the duration of unvoiced phonemes was much longer than that of the voiced counterparts. She observed that on average /t/ took significantly longer than /d/, /p/ took longer than /b/, and /f/ was measured to last almost twice as long as /v/. The standard many-to-one relationship between phonemes and visemes fails to account for these differences in production.

3.3 Discussion

Phonemes are well defined as the unit of acoustic speech. They have the characteristic that exchanging one phoneme in a sequence for a different phoneme changes the meaning of the utterance. The same cannot be said for visemes, which have long been assumed to be the unit of visual speech.

Traditionally, visemes are defined as clusters of visually confused phonemes, where each cluster is associated with a static facial pose. Although there is some overlap between mappings determined by different studies, no two unequivocally agree regarding the number and composition of the viseme groups. This suggests that the relationship between phonemes and visemes is more complex than a many-to-one mapping. This is supported in Figure 2.6 where it is shown that the same phoneme is expressed with a variety of poses in real speech.

Phoneme-to-viseme mappings vary across different speakers, languages and contexts. The majority of the mappings from literature provide only a sparse coverage of the phonemes, as vowels are deemed difficult to cluster and the nature of the methods often restrict the stimuli to contain only a subset of the phonemes. Static visemes are unable to model the important information embedded in the dynamics of speech, and, to date, no mapping considers long term coarticulation effects. As

it stands, visemes are poorly defined and based on a flawed assumption regarding the relationship between acoustic and visual speech.

Chapter 4

Speech Animation

The goal of speech animation is to present the correct articulatory dynamics to synchronise with acoustic speech on a chosen face model. Realistic computer facial animation is challenging as humans are sensitive to the smallest discrepancies from normal behaviour. For speech animation to appear plausible, it needs to be smooth and synchronous to the audio track, address coarticulation and respect the dynamics of the face. It is important that this is done correctly, as poorly animated faces can interfere with the comprehension of the audio speech [110].

One of the earliest forms of character animation involved displaying hand-drawn images in quick succession to create what appears to be a moving image. Typically, moving characters were artistically drawn onto a transparent cel, which was placed over a stationary background image. This method was behind such classics as *Popeye the Sailor* and *Betty Boop* in the early 1930s, and Walt Disney's *Steamboat Willie*, *Snow White and the Seven Dwarves* and many more. Nowadays, this technique is barely used as it is slow and costly, and it requires highly skilled artists.

In the 1970s Parke [120] pioneered the field of computer generated facial animation with his 3D face model which was composed of a few hundred polygons. For animation, poses were keyframed onto the face and the intermediate frames were generated using cosine interpolation. Since then, a vast number of approaches have been developed.

Speech animation techniques can be broadly categorised into three main methods: 1) text-driven methods, which use phoneme labels to select the correct pose or series of poses which are then concatenated or interpolated [15, 24, 27, 33, 38, 45, 46, 100, 101, 124], 2) audio-driven methods, which use the parameterised acoustic speech to estimate the facial pose [13, 36, 42, 144], and 3) performance-driven methods, which map the movement of a speaker onto a facial model [26, 118, 150, 153, 156]. The resulting animation can be rendered on image-based [15] or graphics-based models [27], or a hybrid of the two [144]. Image-based methods typically offer a higher level of realism than graphics-based methods, but are far less flexible in terms of speaker identity and pose.

This chapter provides an overview of the techniques commonly used for speech animation.

4.1 Keyframe Interpolation

Keyframe interpolation is the technique that is most widely used in industry due to its simplicity. It concerns discretising speech into a string of phonemic targets, which are then mapped to viseme targets using a simple lookup table. Given these static targets, an interpolation function is used to generate animated sequences by computing the in-between frames. The interpolation function is based either on static targets [34, 46] or a more complex function that attempts to model coarticulation [27, 38, 45, 124].

Ezzat et al. [46] used a morphing approach to generate the intermediate frames between static visemes using optical flow vectors. While morphing produces a smooth transition, the resulting animation appears over-articulated as the effects of coarticulation are ignored.

Cohen and Massaro [27] extended Löfqvist's gestural production model [95], where the interpolation function is based on exponentially decaying dominance functions and hand-crafted. Each phoneme is modelled with a set of control parameters,

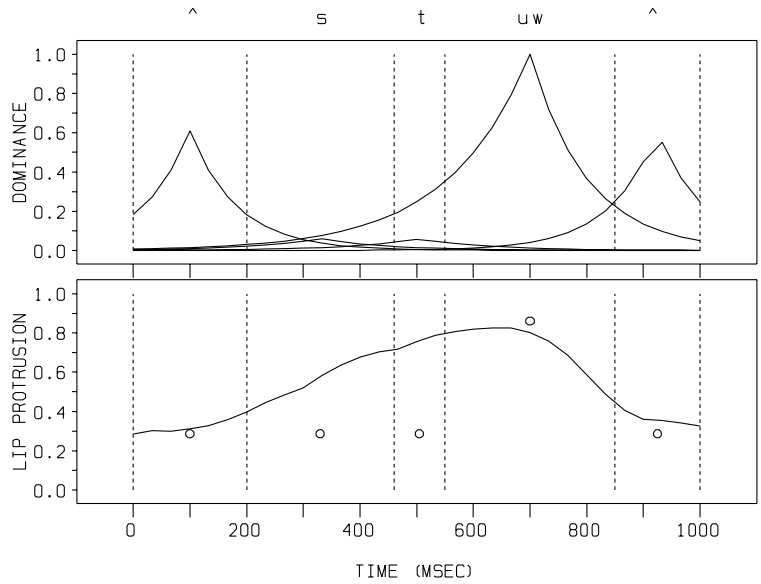


Figure 4.1: An example of Cohen and Massaro's dominance functions for the word "stew" (top) and the resulting value of the lip protrusion control parameter (bottom) taken from [27] with permission.

such as the upper lip position, lip width and degree of jaw rotation. A target value and a pair of negative exponential functions describing the onset and offset dominance is specified for each parameter. The shape of the dominance function is controlled by the duration from the centre of the phoneme segment. Figure 4.1 (taken from [27]) illustrates an example of the dominance functions (top) and the resulting value of the lip protrusion control parameter (bottom) for the word "stew". As the /s/ and /t/ segments have low dominance with respect to protrusion compared with the /u/ segment which has high, wide spread dominance, the onset of lip protrusion occurs early, during the production of /s/.

A similar approach was proposed by Dey et al. [38], in which the influence of each viseme, rather than each phoneme, was modelled with a dominance function. A limitation of these methods is that they fail to ensure that certain targets are realised, such as the closure for a bilabial.

Lazalde and Maddock [89] modelled each viseme as a distribution around an ideal lip pose. To generate speech, a trajectory that attempts to pass through the viseme centroids is optimised with three constraints: 1) boundary constraints which ensure

that the trajectory starts and ends at a particular state with zero acceleration and velocity, 2) range constraints to restrict deviation from the viseme centre, and 3) acceleration/ deceleration constraints so that the mouth moves at a realistic rate.

Pelachaud [124] presented an expressive talking head using Ekman and Friesen's Facial Action Coding System (FACS) [41] units to model expression. Each phoneme was represented by a set of target poses, one for each expression, and the interpolation function was controlled by the dominance of the visemes and the contraction and relaxation times of the muscles.

A disadvantage of the methods described in [27, 38, 124] is that the interpolation function is hand-tuned, which is an adhoc and extremely time consuming process. In [45] and [89] the interpolation function is generated automatically by learning the distribution of the visual features belonging to each phoneme. The extent of coarticulation is implicitly modelled based on the magnitude of a phoneme's variance.

The advantage of keyframe techniques is the relative ease and speed at which animation can be produced. However, these methods typically produce unrealistic results as the natural facial dynamics are ignored in place of an interpolation function. If the interpolation function is defined incorrectly, the animation risks appearing over-articulated, under-articulated or unnatural. Typically, it is also the role of the interpolation function to model coarticulation[1], which is an incredibly complex function, and has yet to be emphatically defined.

4.2 Concatenative Synthesis

Concatenative approaches are widely used in acoustic speech synthesis. Rather than interpolating between static targets for visual speech animation, sequences of speech based on some animation unit are stitched together [15, 24, 33, 100, 101]. The units are typically selected from a training corpus by minimising a cost function that

[1]This is not the case in [34] in which animation is based on static, context-dependent viseme targets.

Figure 4.2: Image-based concatenative synthesis using Bregler et al.'s triphone units. The images are taken from page 6 of [15] with permission.

trades off a measure of similarity between candidate and desired phonetic contexts and the smoothness at the concatenation boundaries.

The units selected from the corpus might be fixed length [15, 100], or variable length [24, 33, 101]. Bregler et al. [15] presented a method based on fixed-length triphones, which are sequences of three phonemes. A dynamic programming algorithm was used to search the training data for the optimal sequence of triphones based on a phoneme-context cost and the distance between overlapping lip shapes. If an example from the training data for a particular phoneme in context was not found, a different phoneme from the same viseme group was substituted at an additional cost. The triphone sequences of the jaw region were retimed, overlapped, cross-faded and stitched onto a background image to generate novel speech animation (see Figure 4.2).

A similar method is described by Ma et al. [100], who introduced the diviseme as a concatenative unit, defined as a transition from one static viseme to another. For novel speech, a sequence of diviseme examples was selected using the Viterbi algorithm to find the best path through a directed motion graph in which nodes represent diviseme instances and edges are weighted with the join cost. For animation, the divisemes were resampled to the correct duration, overlapped and blended.

The disadvantage of using fixed-length segments such as divisemes and triphones is that they fail to account for coarticulation which spans further than two or three units respectively. This prompted Ma et al. [101] to extend the diviseme approach to a variable length animation unit. In their work, all of the viseme strings in the

training data were modelled by a graph, where the nodes represent viseme symbols and the edges were weighted with the concatenation cost. The cost of joining two nodes that appear consecutively in the training data is zero, as they transition naturally with no discontinuity. The Viterbi algorithm was used to determine the optimal path of nodes based on the concatenation cost, and the resulting trajectory is smoothed.

A variable-length unit was also proposed by Cosatto et al. [33]. For animation, the training data was searched for the best matching examples of the phonetic context, producing a set of candidate poses at each frame. These candidates formed a directed graph which were connected by edges which were weighted based on the transition cost. As in [100, 101], the Viterbi algorithm was used to calculate the smoothest path through the graph. If there was a large transition cost between two consecutive poses, the images were blended to prevent jerky movements. Animation with variable-length units appears smoother than with fixed-length units as the longest possible sequences of real data are extracted from the corpus and so there are fewer concatenation discontinuities.

A dynamic animation segment, referred to as an *Anime*, was presented in [24]. An Anime was initially defined as the motion corresponding to a phoneme. However, to reduce the number of motion fragments, the Animes were clustered on their visual similarity, such that each Anime represents a motion that corresponds to a set of phonemes. The Animes form a directed graph with edges modelling the transitions between phonemes in the training data. For novel speech, the search algorithm finds the longest matching sequence of phonemes in the Anime graph. If there were two paths of equal length, the audio features were then compared to determine the best sequence. The animated speech was mapped onto a 3D model. However, as only the facial geometry was captured, the information from the appearance of the speaker, such as teeth visibility and tongue position, is lost and the speech appears under-articulated.

For animation, concatenative methods are somewhat inflexible, as the identity of

the speaker and the range of facial poses are limited to those from the training data. Some attempts at mapping speech to a different face model have been described, for example using radial basis functions [101], however the results are unconvincing.

The advantage of concatenative approaches is that the dynamics of the original speaker are preserved as segments of real video are simply reordered to generate new speech. However, the quality of the speech animation is highly dependent on the amount of data available and the coverage of the phonemes in varying contexts, as finding a particular speech segment in the correct context is key to the process.

4.3 Motion Transfer

Motion transfer concerns mapping real motion data from a talker to a model. This has the advantage of capturing the liveliness and subtleties of facial gestures produced by the performer. Facial motion can be captured using vision-based tracking algorithms, or other marker-based motion capture equipment. For example, Williams [156] tracked a set of markers that were positioned on a speaker's face and then directly mapped them to a scanned 3D model.

This approach can be taken further by using bilinear or multilinear models to separate identity, speech and expression such that the characteristics of the transferred speech can be manipulated so speech can be presented in different emotional contexts or on different faces [26, 118, 150, 153].

A problem with the motion transfer approach is that the mapping from the actor's movements to the deformations on the animated character is often non-linear and can be somewhat complex, especially for non-human characters with different proportions to the speaker. It is also difficult to constrain the lip shapes and motions to a set that are valid for a particular character. To overcome this, Kouadio et al. [83] introduced a motion transfer method where the animated expression was calculated as a weighted sum of basis expressions. A least squares regression is used to learn the weights from the actor's expression.

Figure 4.3: Model-based animation using Brand's *Voice Puppetry*. The images are taken from page 7 of [13] with permission.

Whilst performance-driven approaches are effective for generating realistic animation, motion transfer lacks the flexibility of true synthesis in that the models do not generate the animation data, rather an actor is always required.

4.4 Model-based Synthesis

A more flexible approach is to use generative models to synthesise animation parameter trajectories. Statistical models can be employed within a probabilistic framework to model the joint distribution of acoustic and visual speech [13, 36, 42, 144], exploiting the correlation between the visual and auditory components of speech. Given novel acoustic speech, these distributions can be sampled to estimate the maximum likelihood facial animation parameters, which can then be applied to the visual model.

Brand's *Voice Puppetry* [13] modelled the position and velocity of the facial features and the relationship to the audio features. An acoustic HMM estimates the state sequence given a novel audio track, and an HMM trained on the visual features drives the animation using the estimated state sequence. A geodesic interpolation technique is used on the facial motion parameters to ensure smoothness. A sequence of animation generated using Voice Puppetry is shown in Figure 4.3 [13]. The benefit of using an HMM is that it can account for context across an entire utterance, regardless of its length. However, the resulting animations appear overly smoothed and under-articulated.

Theobald and Wilkinson [144] used canonical correlation analysis (CCA) to model the linear relationship between audio and visual speech by finding a set of audio and visual basis vectors which maximise the correlation between the two sets of features. Linear regression was then used to find the transformation that maps the audio parameters (projected on the basis) to the visual parameters. Coarticulation was modelled by appending features to the right and left of each frame. However, the authors note there were instances where mapping from auditory to visual parameters in this way did not perform well. For example, sounds formed towards the back of the mouth can appear very different visually, and it is difficult to model the relationship between these audio and visual features using a simple linear model.

Deena and Galata [35] used a Gaussian process latent variable model framework to learn a shared latent space between the audio and visual speech. This model can be viewed as a non-linear extension of CCA. A limitation of this approach is that all of the phonemes share a single model, irrespective of the variability in dynamics. To address this, the model was further extended by augmenting it with switching states that represent the varying dynamics of speech [36]. Each switching state, which represents a commonly occurring sequence of phonemes, is modelled with a shared Gaussian process dynamical model. New speech animation is generated by inferring the state for each audio frame and calculating the latent variables from the shared model. The visual features are then calculated from the latent variables.

Model-based methods have the advantage that the facial poses are learned from observed data and they do not require the amount of data that is necessary for concatenative techniques. However, they are limited to the speech model which they are trained.

4.5 Discussion

For believable speech animation, it is necessary that the simulated motion is smooth and reflects the dynamics and coarticulation effects of real speech. Humans are finely

tuned to the subtleties of facial movement, making it a challenging task.

The simplest technique for generating animated speech is keyframe interpolation. The advantage of this method is that, to animate any model, the animator simply defines one pose for each phoneme or viseme. However, the natural dynamics of the speaker are ignored as it is the role of an interpolation function to smoothly transition across poses and to simulate coarticulation. Not enough is known about coarticulation for this method to produce realistic results.

Concatenative methods involve stitching together segments of real speech to produce animation. They therefore maintain the original dynamics of the speaker. Coarticulation can be modelled by minimising a cost function across an entire utterance. However, this method requires a large amount of training data, and the synthesised motion is not easily mapped to different facial models.

Motion transfer and model-based approaches are also limited to models that have a similar geometry to the speaker. Motion transfer yields high quality animation, as the motion is mapped directly from the speaker but is inflexible as an actor is always required. Model-based methods are more flexible as they are able to generate novel speech by sampling from a probabilistic distribution which models the relationship between audio and visual speech information.

In the remainder of this thesis a novel, concatenative approach for speech animation is described, that goes some way towards overcoming the limitations of previous approaches. This unit maintains the natural dynamics of the visual speech, and it can be applied to any form of graphics model.

Chapter 5

Data Capture and Visual Speech Modelling

This chapter first describes the capture and annotation of an audio-visual database, and reviews methods for video tracking and parameterising facial features. The choice of features used in this work is then discussed, and tracking and parameterisation using active appearance models is described.

For visual speech analysis, a database containing synchronous audio and visual speech is required. The content needs to be carefully considered, as the video quality and the nature of stimuli are critical to the outcome of the analysis. From a recorded database, a visual parameterisation is derived by locating the oral region in each of the movie frames and extracting information regarding the geometry and appearance of the speech-related facial features. It is important that the area of the face described by this parameterisation sufficiently captures the speech-related movements, and that the features are discriminative enough to distinguish between the subtleties of speech gestures.

5.1 Audio-Visual Speech Database

The specification of the audio-visual speech database is important as it is has direct impact on the outcome of the work described in this thesis. The image resolution needs to be sufficiently high to ensure that the extracted features capture fine detail that provide discriminative speech information. It is also necessary that the frame rate is sufficient to capture the dynamics of the speech, including plosives which are produced using rapid motion. To enable analysis of longer term coarticulation effects, the stimuli should be in the form of continuous speech and they should contain a good coverage of phonemes in different contexts. It is also necessary that the speaker's face is appropriately positioned within the video frame so the speech articulators are clearly visible, and that the face is illuminated such that there are no intrusive shadows.

There are a number of publicly available audio-visual speech databases, such as CUAVE [122], AV Letters [107], XM2VTSDB [112] and GRID [29]. However, all are restricted to isolated words [107, 122] or limited vocabulary [29, 112], and are therefore unsuitable for use in this work.

Preliminary analysis was initially performed using the LIPS2008 database [143], which contains 278 phonetically balanced sentences spoken by a female speaker. However, it soon became apparent that more speech was required to ensure a complete coverage of speech gestures. Thus, a larger database was recorded, which was entitled KB-2k. KB-2k contains an actor reciting the 2342 sentences from the TIMIT sentence list [117], plus an extra repetition of 200 of these sentences. The actor read the sentences from a teleprompter in an American English accent and maintained a neutral speaking style throughout the recording .

The TIMIT sentence list is composed of 2 sentences that were designed to expose variation in dialect, 450 phonetically-compact sentences and 1890 phonetically-diverse sentences taken from the Brown corpus [54] and the Playwright's Dialog [68]. The phonetically-compact sentences were designed to provide good coverage of pairs

of phones, with extra occurrences of phonetic contexts that were thought to be either difficult or of particular interest [117]. The phonetically-diverse sentences were selected to maximise the variety of allophonic contexts in the text and were designed to add diversity in terms of sentence type and phonetic context.

The video was recorded at 29.97 frames per second at a resolution of 1920 by 1080 progressive scan and it runs to approximately 8 hours. Both a frontal view and side view of the actor were captured[1]. A lighting rig was positioned to illuminate the face such that the tongue and teeth were visible, and that the view of the face was not impeded by shadows. The recording process was supervised, and if the actor mispronounced a sentence he was asked to say it again and the outtakes were discarded. A selection of frames from the database is shown in Figure 5.1.

The audio speech was captured at a sampling rate of 48 kHz via two microphones; one onboard the camera, and a tie-clip microphone attached to the actor's shirt. The sentences were phonetically segmented using the audio, and they were annotated using the ARPAbet phonetic notation code. Annotation was performed manually by five students from Carnegie Mellon University, all of whom had completed a course on phonetics. The labels were then checked and adjusted so that they were error free and consistent across the entire database.

5.2 Parameterising Visual Speech

There is an enormous quantity of data in video sequences, as each frame contains millions of pixels that are vastly redundant and non-speech related. From this high dimensional data, the challenge is to extract a set of low-dimensional feature vectors that contain visual speech information with good discriminatory power. That is, for each movie frame, an accurate representation of the configuration of the visible speech articulators is required. The initial stage of this process involves determining the location of the lips and jaw from which features can then be extracted from

[1]Only the frontal view was used for the work described in this thesis.

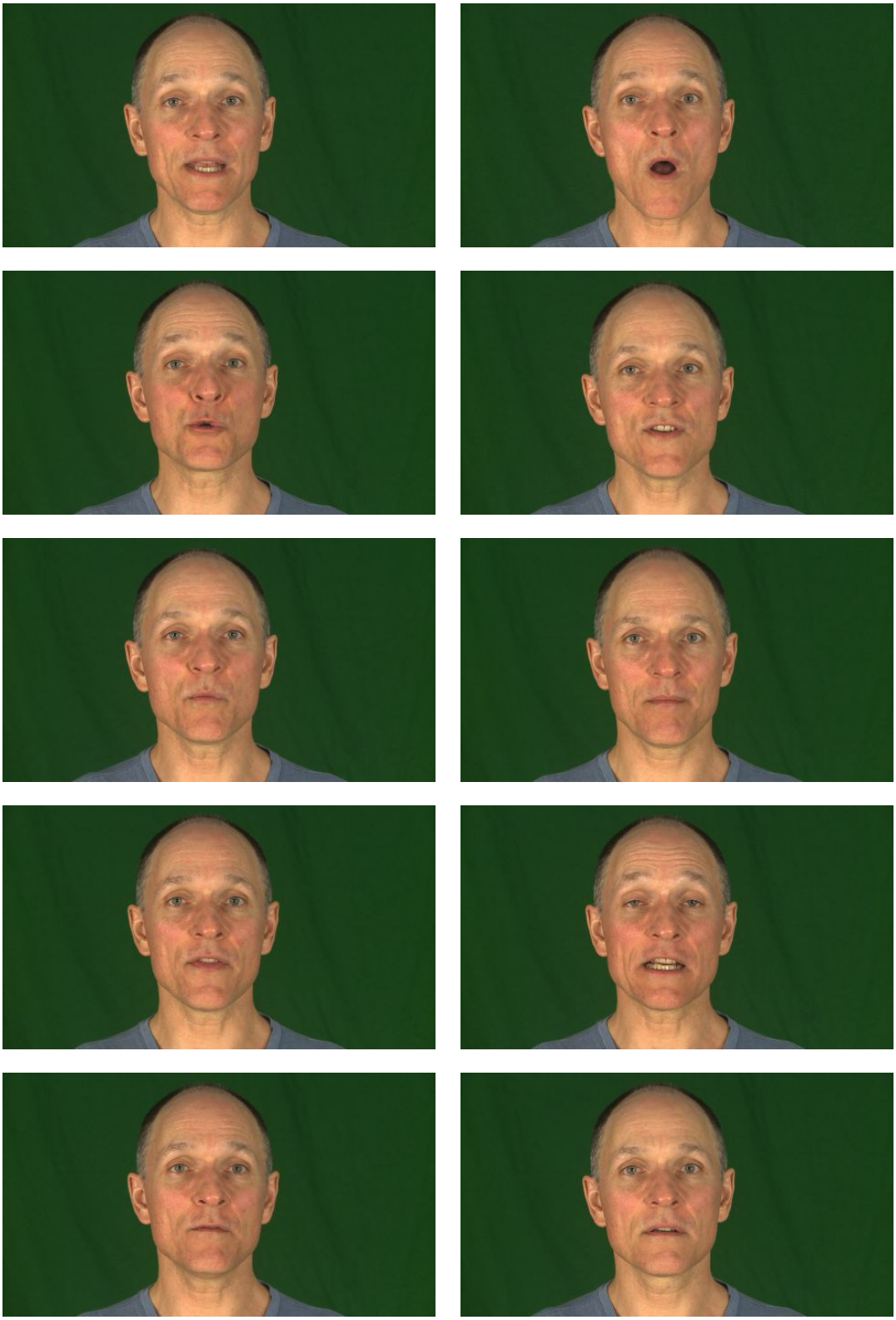


Figure 5.1: A selection of frames extracted from the KB-2k database illustrating the illumination conditions and a sample of the actor's poses.

the image. These features might describe the shape or appearance of the visible articulators, or they might represent other information, such as the direction or magnitude of the articulator motion.

Feature extraction for acoustic speech is well established. Audio is captured using a microphone and can be encoded using Linear Predictive Coding (LPC), Mel-Frequency Cepstral Coefficients (MFCCs), Line Spectral Frequencies (LSF) or Formants [57]. However, no such standards exist for the visual modality and the choice of visual features remains somewhat arbitrary.

Robust feature extraction for faces is a difficult problem due to the large appearance variation between and within subjects. Between-subject variation accounts for differences in skin tone, facial hair and facial proportions, whereas within-subject variation accounts for the changes in appearance caused by speech, expression and head pose. Effective facial tracking is highly dependent on the environmental conditions, as vision-based algorithms rely on predictable or constant global lighting. This is because the colour, degree and direction of illumination alters the appearance of the face. As shadows form, pixel intensities change. For visual speech analysis, it is necessary that the facial parameterisation represents the speech information independent of the variation in illumination. For this reason, speech is typically captured under controlled conditions, whereby expression, head motion and illumination are fixed. Some methods depend upon lipstick to make the lips and skin-tone more contrastive. Others use stick-on reflective markers or head-mounted cameras. However, these methods potentially interfere with the naturalness of the speech as they are intrusive and might affect the way the talker speaks.

Methods for tracking and parameterising facial features can be broadly grouped into shape-based and image-based approaches. This section describes methods of facial feature extraction, and presents a discussion on the type of features that are important for analysis.

5.2.1 Shape-based Methods

The shape of the visible articulators can be parameterised with simple geometrical measurements, such as lip height, width and roundedness. For example, Brooke and Templeton [20] thresholded images of the oral region and performed a valley edge detection algorithm, before manually extracting the mouth height, width and the size of the inner mouth area for clustering into viseme classes. Goldschen et al. [58] hand-segmented the region of interest (ROI) surrounding the lips, and then automatically extracted features describing lip rounding, height, width and perimeter, and the size of the inner lip area. However, manual tracking is incredibly slow, and prone to human error. An automated approach was described by Petajan et al. [127] who used a pattern matching algorithm to track the nostrils in thresholded images of the face. The oral region was assumed to appear at a fixed distance from the nostrils. Although it worked successfully for their data, this method is extremely sensitive to speaker, scale and head pose.

A more liberal representation of the lip position can be extracted by modelling the x and y coordinates of fiducial landmarks positioned on the contour. There are a number of edge-detection algorithms for extracting the shape of the lip contour. For example, Snakes (also known as Active Contour Models) [79] model the contour of an object with a spline, represented as a piecewise polynomial curve. The polynomials are constrained such that the curves meet at the join with continuous first and second order derivatives. To fit to an image, the coefficients are automatically adapted to minimise an energy function that is a combination of internal energy, which controls the flexibility of the curve, and image energy, which pulls the snake towards the edges.

A problem with this approach is that the shape is unconstrained, so there is no certainty that the fitted shape will be valid. Human lips are geometrically complex, but the shape varies with a number of distinct degrees of freedom constrained by the physiology of the face [18]. Point distribution models (PDMs) provide a method of extracting these degrees of freedom by capturing the mean shape and principal

modes of variation from labelled training data. PDMs provide a method of statistically modelling the shape, so that constraints can be enforced so only valid contours are produced.

To learn a PDM, a set of training images are typically hand-labelled with landmarks describing the lip contours and other salient facial locations. Once labelled, the landmarks are aligned to normalise for scale, rotation and translation and then transformed using principal components analysis (PCA). PCA is detailed in Appendix A. A particular configuration of the lips can then be represented as a point in feature space, and the set of all possible (legal) configurations of the lips can be represented as a smooth surface within the learnt manifold. The training images are carefully selected to represent the extreme poses, so that the principal modes of variation represent all valid facial deformations, and the dimensionality of the feature space represents the number of degrees of freedom of the lips.

Bregler and Omohundro [18] used PCA to learn a PDM from the frames that were successfully tracked using the conventional snake algorithm. The lips were then re-tracked by maximising the grey-level gradients along the contour calculated at normals to each of the landmark points. This time, the shape was constrained to lie within the learned shape space.

An alternative approach for fitting a PDM to an image is using active shape models (ASMs) [32]. To fit a PDM to an image using the ASM method described in [32], the profile normals at each of the landmark points are examined, and the shape boundary is shifted towards the strongest edge. The model parameters are then updated to reflect the new shape, while constraining the shape to lie within the learned manifold. The process is then iterated until the change in shape is sufficiently small. The problem with this approach is that contour gradients are often unreliable for defining the lip boundary as they are affected by conflicting information, such as shadows and facial hair, and vary depending on the position of the lips. A good example of this can be seen in [99]. As the lips are set against the similar flesh tones of the surrounding skin, the gradients can be a poor basis

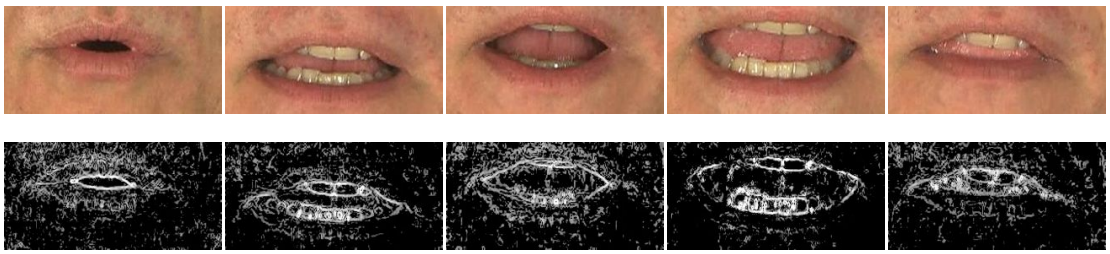


Figure 5.2: The gradient magnitude of the lip region for a selection of frames from the KB-2k dataset. As the lips are set against the flesh tones of the skin, the contours are often barely visible, or segmented. The gradients therefore function as a poor basis for tracking.

for segmentation, especially along the lower outer lip (see Figure 5.2). Luettin [98] modified this approach by modelling the pixel intensities along the profile normals for each of the landmark points, and stacking them to construct a global profile vector. The global profile mean and covariance was then calculated and the modes of variation were learnt using PCA. For each frame, the model parameters were iteratively updated, and the mean squared error between the model and the image profile values was calculated until convergence. They adopted a simplex method for shifting the shape towards a minima of the cost function.

The main limitation of ASMs is that only the image information that is local to the shape boundary is used to fit the model to a new image, so the model is prone to losing track of the contour. The next sections discuss methods that use the full appearance of the oral region.

5.2.2 Image-based Methods

Image-based methods involve processing the pixel intensities of the image, usually within a region of interest (ROI) surrounding the lips, and sometimes the cheek and jaw [14]. Direct analysis of the raw pixel intensities has the disadvantage of high dimensionality, and high data redundancy as neighbouring pixels are likely to change at a similar rate. Therefore, dimensionality reduction algorithms are typically employed to make the visual features more discriminant. These meth-

ods include processing the pixel intensities with a discrete cosine transformation (DCT) [88, 97], discrete Fourier transform (DFT) [16, 40], linear discriminant analysis (LDA) [40, 97] or principal components analysis (PCA) [17, 40], and retaining only the features that account for the low frequency information, or a large amount of variation of the original image. These techniques have an additional benefit in that they produce generative features that are generally less sensitive to noise.

For audio-visual speech recognition, Lucey et al. [97] extracted a rectangular, 32×32 pixel ROI encompassing the lips by locating the eyes and nose, and then using a classifier to locate the lip corners. A 2D DCT was then applied to the mean-subtracted pixels within the ROI, and the high frequency information was ignored. Finally, LDA was performed to further reduce the dimensionality of the features. Similarly, Bregler and Konig [17] first tracked the contour of the lips using a snake-like algorithm, and then extracted a rectangular, 24×16 pixel ROI which was centred on the lips. The appearance of the lips was modelled by applying PCA to the scaled ROI. Inspired by Turk and Pentland's Eigenfaces [146], this approach was referred to as Eigenlips.

The problem associated with ROI-based appearance modelling is that it captures variation due to both shape and appearance, as the change in intensity over a pixel is captured, rather than the change over a particular location on the lips. Instead, it is preferable that each pixel represents the same feature on the face, thus allowing separation of appearance and shape information. Active appearance models [30] provide a means for accomplishing this, and are described in the following section.

5.2.3 Active Appearance Models

Active Appearance Models (AAMs) [30] are a compact, generative, statistical representation of *both* the shape and the appearance variation in a set of images. The shape of an AAM is defined by the two-dimensional vertex locations of a mesh that

delineates the contours of the visible articulators:

$$\mathbf{s} = \{x_1, y_1, x_2, y_2, ..., x_N, y_N\}^T,$$

where (x_i, y_i) are the coordinates of the i^{th} landmark on the image.

Usually a model is built by first hand-labelling a set of training images with the vertices that define a triangulated mesh. To capture only speech-related variation, care is taken to place each of the landmarks accurately and consistently around the contours of the facial features at regular intervals. The training meshes are normalised for translation, scale and rotation by solving for the similarity parameters to align the shapes. In particular, using:

$$\mathbf{s} = M(\omega, \theta)[\mathbf{s} - \mathbf{t}], \tag{5.1}$$

A generalised Procrustes analysis [59] can be used where $\mathbf{t}$ describes the translation to zero centre the landmarks,

$$\mathbf{t} = \{t_{x_1}, t_{y_1}, t_{x_2}, t_{y_2}, ..., t_{x_N}, t_{y_N}\}^T, \tag{5.2}$$

and $M(\omega, \theta)$ describes scaling of ω and rotation of θ. To align all of the training images, all of the shapes are first aligned and then the mean of the aligned shapes is calculated, and all of the shapes are realigned to the mean. This is repeated until convergence [32].

From the aligned shapes, the mean and covariance matrix is calculated and PCA is applied to obtain the eigenvectors and eigenvalues of the covariance matrix. This provides a compact representation of a shape in the form:

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{m} \mathbf{s}_i p_i, \tag{5.3}$$

where $\mathbf{s}_0$ is the mean shape and the vectors $\mathbf{s}_i$ are the eigenvectors of the covariance matrix corresponding to the m largest eigenvalues, which represent the most signif-

icant modes of variation. The coefficients p_i are the shape parameters, which define the contribution of each mode in the encoding of $\mathbf{s}$. This model describes the legal shape deformations learnt from the training examples, and any valid shape can be approximated with the parameters $\mathbf{p}$. Conversely, given a set of example points $\mathbf{s}$, the parameters $\mathbf{p}$ can be calculated using:

$$\mathbf{p} = \sum_{i=1}^{m} \mathbf{s}_i^T (\mathbf{s} - \mathbf{s}_0). \tag{5.4}$$

The appearance of an AAM is defined over the pixels within the triangulated mesh formed from the hand-labelled landmarks, $\mathbf{x} = (x, y)^T \in \mathbf{s}_0$. To construct the appearance, each of the training images are first warped to the mean shape, $\mathbf{s}_0$. PCA is then applied to the shape-normalised images to give a compact model of appearance variation of the form:

$$A(\mathbf{x}) = A_0(\mathbf{x}) + \sum_{i=1}^{n} \lambda_i A_i(\mathbf{x}) \qquad \forall \mathbf{x} \in \mathbf{s}_0, \tag{5.5}$$

where the coefficients λ_i are the appearance parameters, $A_0(\mathbf{x})$ is the base (mean) appearance, and the appearance images, $A_i(\mathbf{x})$, are the eigenvectors corresponding to the n largest eigenvalues of the covariance matrix.

It is intuitive that the shape and appearance features are somewhat correlated, as they are synchronous to the acoustic speech. To decorrelate the two sets of features, the shape and appearance parameters are stacked, and a third PCA is applied, giving:

$$\mathbf{b} = \begin{pmatrix} w\mathbf{p} \\ \boldsymbol{\lambda} \end{pmatrix} = \sum_{i=1}^{q} \mathbf{j}_i c_i, \tag{5.6}$$

where $\mathbf{p}$ is a vector of shape parameters, $\boldsymbol{\lambda}$ is a vector of appearance parameters, $\mathbf{j}_i$ are the basis vectors spanning the combined shape and appearance space and $\mathbf{c}$ are the parameters that describe the combined shape and appearance variation of the

lips and jaw. The coefficient w normalises for the energy:

$$w = \sqrt{\frac{\sum_{i=1}^{n} \sigma_{\lambda_i}^2}{\sum_{i=1}^{m} \sigma_{p_i}^2}}, \tag{5.7}$$

where m and n are the dimension of the shape and appearance parameters respectively, and $\sigma_{p_i}^2$ and $\sigma_{\lambda_i}^2$ represent the variance captured by each dimension of the respective model.

Fitting an AAM to a face can be performed using a wealth of algorithms which typically involve minimising the difference between the face image and a synthesised image from the appearance model. A good overview of fitting algorithms for AAMs can be found in [31].

5.2.4 Selecting Features for Visual Speech Analysis

It is intuitive that the appearance of the lips, teeth and tongue are all important for decoding visual speech, and so analysis is typically performed on features that describe the oral region [17, 97, 99]. However, other speech information, for example lip protrusion and cheek puffing, is embedded in the shadows and creases that appear over the entire lower face area. This was confirmed by Potamianos and Neti [128] who measured a significant increase in accuracy of an audio-visual speech recogniser when the jaw and cheek region were included in the ROI, over using the lip-only area. This extra information is also beneficial for humans, as Ijsseldijk [70] measured a lower lip-reading accuracy when subjects were presented with only the lip region of the face, and Scheinberg [136] discovered that certain points located on the cheeks and jaw provided the information required for visual discrimination between the phonemes /p/ and /b/. Jiang et al. [75] also found that jaw and cheek information is beneficial over lip-only information when comparing physical and perceptual measurements between visual speech syllables by analysing confusion matrices.

The dynamics of the visual features are often more discriminative than the static

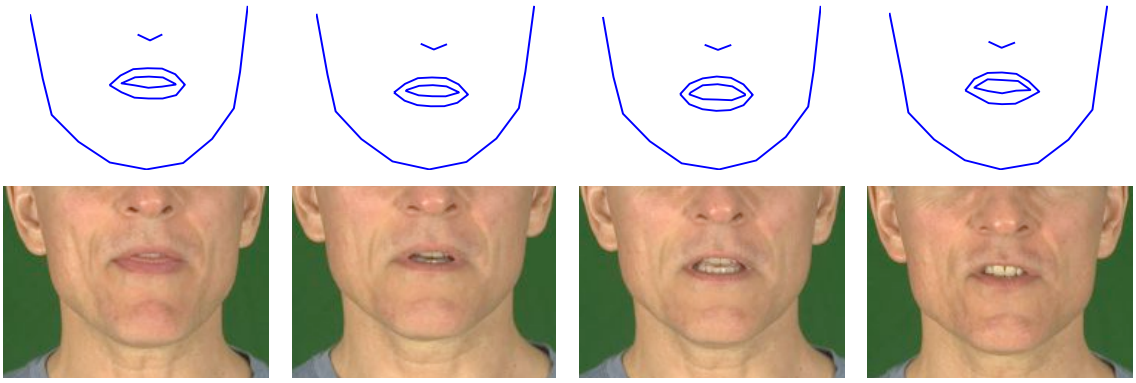


Figure 5.3: Illustrating the importance of appearance information to visual speech analysis. The top row shows the feature boundaries for the corresponding movie frames on the bottom row. It is apparent that appearance information is important for discriminating between visual speech poses.

features. Typically, the dynamics are represented by the first and second derivatives, where the first derivative, $\Delta = \frac{\delta y}{\delta t}$, describes the rate of change and the second derivative, $\Delta\Delta = \frac{\delta^2 y}{\delta t^2}$, describes the acceleration. For example, Bregler and Omohundro [18] measured $\approx 3\%$ improvement on audio-visual word recognition when acceleration features were included, and Lucey [96] discovered that extracting features on difference images, rather than the original images, decreased the word error rate by 10%. Goldschen et al. [58] experimented with a set of static and dynamic features for automatic lip-reading, and concluded that the majority of features that are salient for visual speech recognition pertained to the dynamics. It is also thought that dynamic features are likely to be more robust across speakers, as they encode the rate of change rather than absolute values [99].

The advantage of using shape-based features for visual speech analysis is that they can be transformed such that they are invariant to scale, rotation and translation, and, if the tracking is robust, to illumination. However, they fail to describe important speech-related aspects of the face, such as the presence of teeth, position of the tongue and shadowing caused by lip protrusion. These phenomena play an important role in speech perception, as they help to distinguish between sounds. For example, Figure 5.3 illustrates the feature contours (top) and the image frames (bottom) for a variety of speech poses. It is clear that the shape of the lips is very

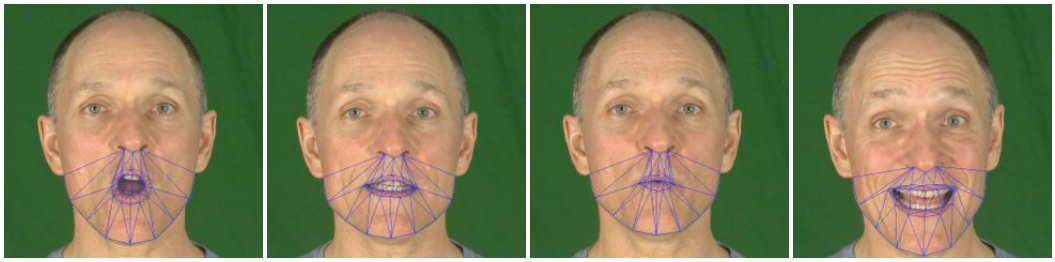


Figure 5.4: A selection of training images used to build the AAM that have been manually annotated with 34 landmarks demarcating the lips and jaw.

similar in each of the examples, but the teeth and tongue are positioned differently, altering the perceived visual meaning. It is therefore apparent that both shape and appearance information is necessary to adequately represent visual speech. To further validate this, the addition of appearance information to shape alone has been found to significantly improve the accuracy for automatic lip-reading [18, 88, 107].

AAMs were chosen to parameterise the visual speech in the KB-2k dataset as they provide a compact statistical representation of both the shape and appearance variation in a set of images. They are also generative, which allows for an *analysis by synthesis* approach to tracking. This is generally more robust and accurate than other approaches, as it is based on the full appearance of the tracked object. AAM features have also been shown to outperform other shape and appearance based features for visual word recognition [88].

5.3 Feature Extraction for KB-2k

An AAM was learnt using 120 training images, each annotated with 34 landmarks, comprising of 12 points demarcating the outer lip, 10 the inner lip, 3 the nostrils and 9 the jaw contours. Figure 5.4 shows the region of the face captured by the AAM.

Figure 5.5 shows the mean and the 11 modes of variation at ± 3 standard deviations about the mean of the shape model. These modes account for 95% of the

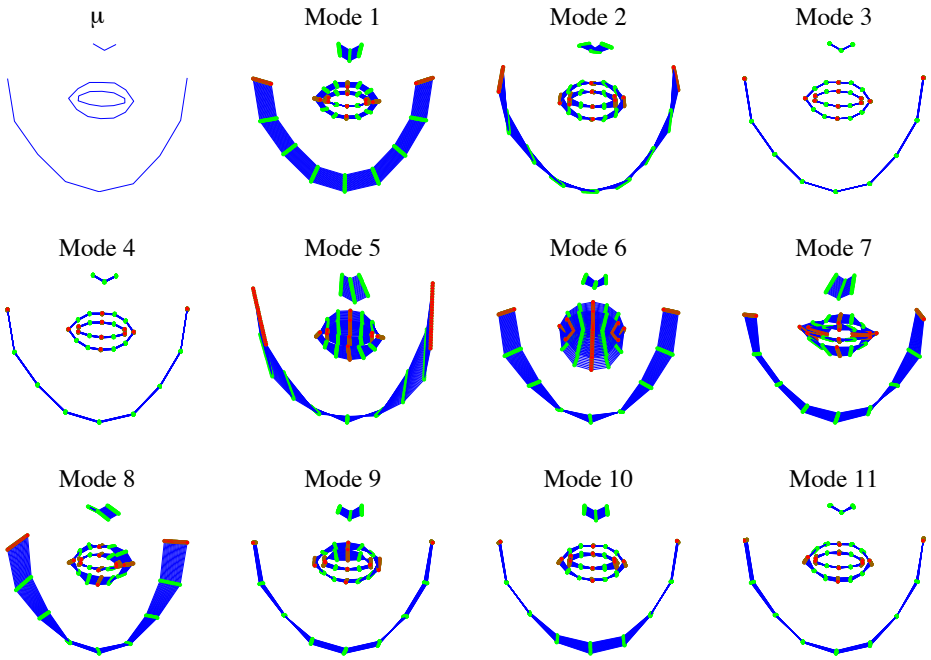


Figure 5.5: Modes of variation for the AAM shape model at ±3 standard deviations about the mean. Modes one to four describe similarity transformations and are discarded for analysis.

overall shape variation in the KB-2k database. The first 4 modes describe scale, rotation and x, y translation variation respectively[2]. Modes 5 and 8 describe nodding and shaking of the head, modes 6 and 7 describe lip opening and rounding and the remaining modes describe more subtle variations.

The mean and first 4 modes of appearance variation for the KB-2k dataset are shown in Figure 5.6. Mode 1 describes the visibility of the teeth, mode 2 describes lip protrusion and spreading, and the remaining modes describe more subtle variations. In total, 95% of the appearance variation is accounted for in 88 modes.

From Figures 5.5 and 5.6 it is apparent that shape mode 7 and appearance mode 2 both appear to encode lip-rounding. To decorrelate the features, a combined model is generated using Equation 5.7, producing an 80 dimensional space which describes the variation in both shape and appearance. Figure 5.7 shows the mean and the

[2]These modes of variation are useful for tracking but are removed for analysis, as they are not meaningful in the sense of speech as they describe the position and orientation of the head.

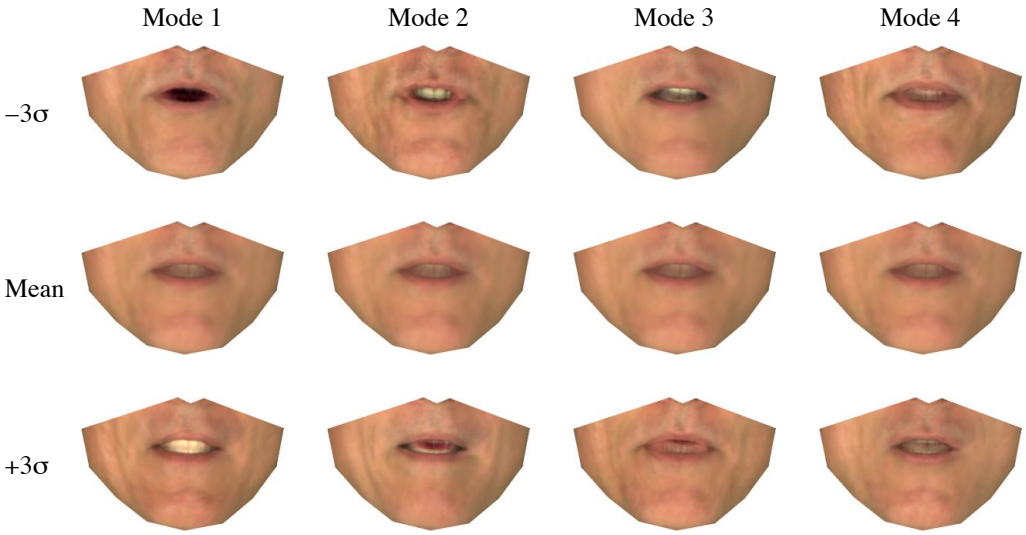


Figure 5.6: The first four of modes of variation for the AAM appearance model at −3 (top) and +3 (bottom) standard deviations about the mean (middle), shown on the mean shape. In total 95% of the overall appearance variation is accounted for in 88 modes.

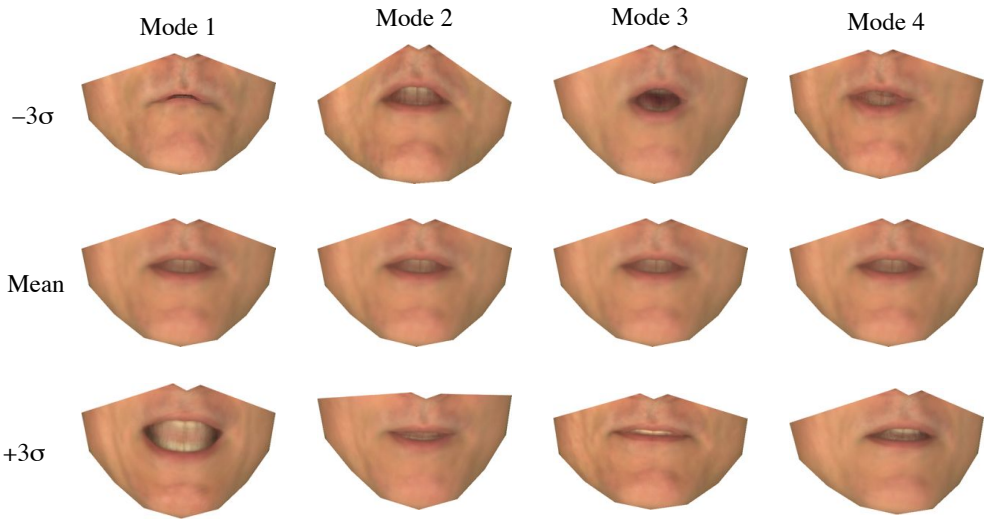


Figure 5.7: First four modes of variation for the combined shape and appearance model at −3 (top) and +3 (bottom) standard deviations about the mean (middle). The combined model contains 80 modes of variation.

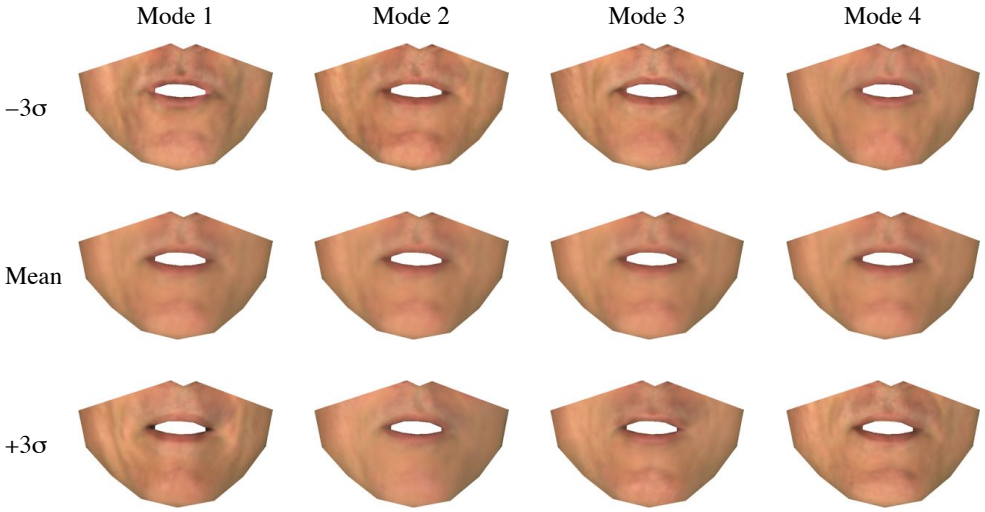
first four modes of variation corresponding to the highest eigenvalues for the combined AAM model.

In this work tracking was performed using the *inverse compositional project-out* algorithm [106]. For each frame, the AAM is initialised on the tracked landmarks of the previous frame. Given the trained AAM, all eight hours of 1080p video were tracked and the speech-related frames were presented for analysis. The tracking was visually inspected, and for any frames where the model lost track the frame was re-fitted.

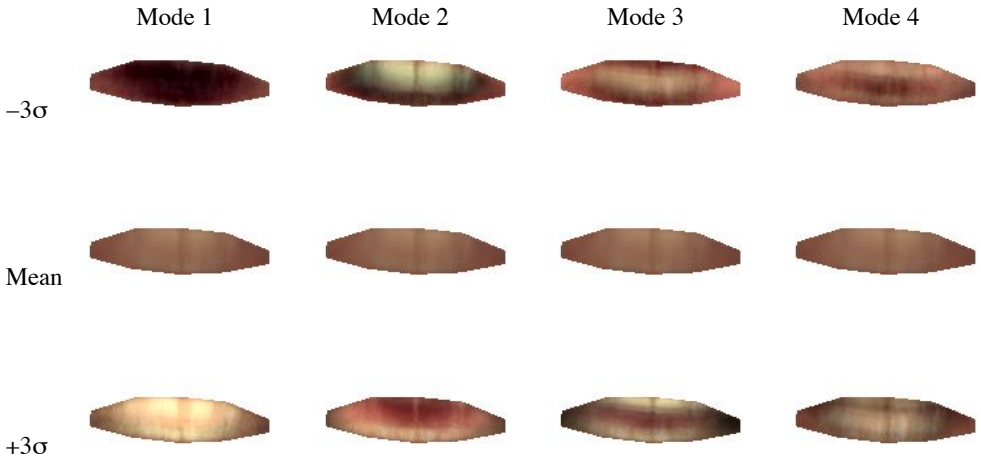
5.3.1 Multi-segment AAMs

Rather than building an AAM with a single appearance component (i.e. performing a single PCA on all of the pixels within the base mesh), instead a multi-segment AAM [145] is built, where different regions of the face are modelled as separate appearance components. This allows greater PCA modelling accuracy of the inner mouth, which is generally not linearly related to the surrounding appearance. A large amount of information regarding the state of the teeth and tongue is conveyed in the appearance of the inner lip area. Thus, a more discriminative set of features is attained by modelling more of the variation in this area [88].

To construct a multi-segment AAM the images are segmented into two sub-regions, one containing the inner-lip area and the other containing the remainder of the lower face pixels. Independent appearance models are then constructed for these sub-regions using Equation 5.5. Figure 5.8 shows the four modes of variation corresponding to the highest eigenvalues for the two appearance models in the multi-segment AAM. The segments are modelled with 46 and 10 modes respectively. The shape parameters, and two sets of appearance parameters are concatenated and



(a) Segment 1: The lips and jaw



(b) Segment 2: The inner mouth

Figure 5.8: Modes of variation for the appearance models of a multi-segment AAM at -3 (top) and $+3$ (bottom) standard deviations about the mean (middle), shown on the mean shape. In total, the segments are modelled with 46 and 10 modes respectively.

normalised and a third PCA is applied, giving:

$$\mathbf{b} = \begin{pmatrix} w_p\mathbf{p} \\ w_\lambda\boldsymbol{\lambda}_1 \\ \boldsymbol{\lambda}_2 \end{pmatrix} = \sum_{i=1}^{q} \mathbf{j}_i c_i \tag{5.8}$$

where $\mathbf{p}$ are the shape parameters, and $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are the appearance parameters for the two segments of the model. The weights, w_p and $w_{\boldsymbol{\lambda}}$ are defined as:

$$w_p = \sqrt{\frac{\sum_{i=1}^{n_2} \sigma^2_{\lambda_{2_i}}}{\sum_{i=1}^{m} \sigma^2_{p_i}}}, \qquad w_\lambda = \sqrt{\frac{\sum_{i=1}^{n_2} \sigma^2_{\lambda_{2_i}}}{\sum_{i=1}^{n_1} \sigma^2_{\lambda_{1_i}}}}, \tag{5.9}$$

where n_1 and n_2 are the number of dimensions corresponding to the first and second appearance components, m is the number of dimensions corresponding to the shape, and $\sigma^2_{\lambda_{1_i}}$, $\sigma^2_{\lambda_{2_i}}$ and $\sigma^2_{p_i}$ represent the variance captured by the i^{th} dimension of the respective model.

This generates a compact, 20-dimensional feature vector, $\mathbf{c}$, which describes the shape and appearance of the lips and jaw at each movie frame. The first three modes of variation of the combined model are shown in Figure 5.9.

5.4 Discussion

In this chapter an audio-visual dataset entitled KB-2k is described containing a speaker uttering eight hours of speech. The stimuli was in the form of sentences containing a good coverage of phonemes in different contexts. The dataset is phonetically segmented manually, and the visible articulators are tracked and parameterised using active appearance models.

AAMs provide a convenient way to model both shape and appearance variation of the visible articulators, both of which are important features for speech analysis. The AAM was constructed by performing independent PCA on the landmarks and pixel intensities from some hand-labelled training images. A third PCA is then applied to

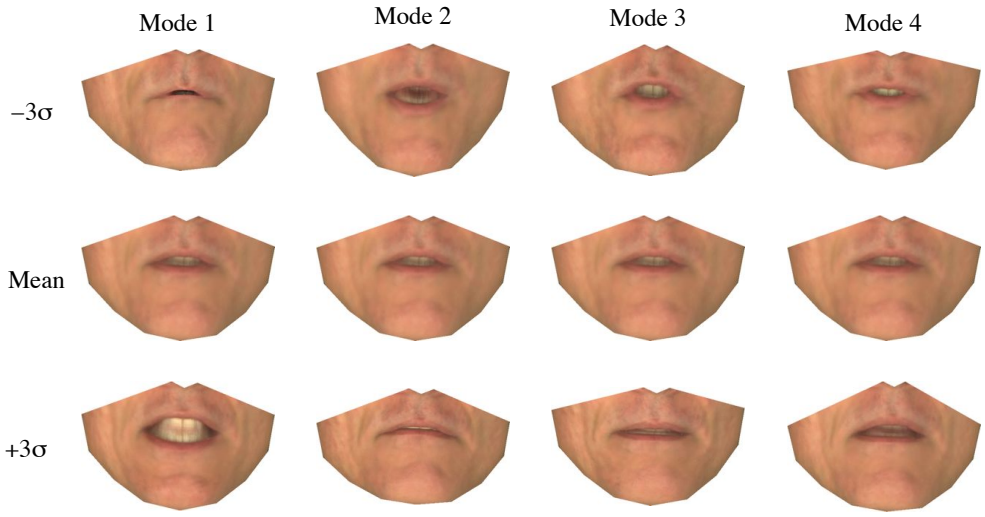


Figure 5.9: Modes of variation for the combined shape and appearance multi-segment model at −3 (top) and +3 (bottom) standard deviations about the mean (middle).

the stacked shape and appearance features to remove correlated information. The model was used to track the jaw and lips in all of the video frames. For analysis, a multi-segment AAM was constructed such that the inner lip area is modelled independently of the rest of the face, meaning that more information regarding the teeth and tongue is described by the features.

The AAM features described in this chapter serve as a basis for the remainder of the work described in this thesis.

Chapter 6

Dynamic Visemes for Speech Animation

This chapter introduces a new *dynamic* unit of visual speech that overcomes some of the issues associated with standard static visemes, which are determined by clustering phonemes. The proposed unit is derived from an analysis of real visual speech, and it is not tied directly to the underlying phones. Rather, a series of canonical speech gestures are determined that, after clustering, form sets of related gestures that are referred to as *dynamic visemes*. These visemes each serve a particular visual function as they represent a specific action on the lips.

This chapter first describes the process for identifying visual speech gestures by segmenting the AAM parameters into short subsequences of movements. An overview of clustering techniques, and the criteria for clustering the identified gestures into groups of related movements is presented. The relationship between the visual and audio units is then described, and finally, the efficacy of dynamic visemes for speech animation is evaluated.

6.1 The Idea Behind Dynamic Visemes

The main drawback of traditional visemes is that the units are based on a visual representation of audio units of speech. Because of this, they do not truly represent the units of visual speech, which is why it is necessary to further post-process animated speech with complicated coarticulation models. Dynamic visemes are different as they are learnt from the visual information. They are determined by first segmenting visual speech into short subsequences of non-overlapping movements, referred to as gestures, each of which describes a short, isolated lip motion. These gestures are then clustered into groups of related gestures, which represent the dynamic visemes. The idea is that the gestures that appear within a cluster have the same visual function, and so replacing one of the gestures with another from within the viseme cluster would not change the visual meaning of the utterance. However, replacing a gesture with one from a different viseme cluster would affect the perceived meaning of the (visual) speech utterance.

The advantage of dynamic visemes is that they describe the finite set of speech related movements of the visible articulators. This means that coarticulation is embedded within the units, and the natural dynamics of the articulators is preserved. Since all gestures that appear within a dynamic viseme class represent the same speech movement, only one example of each cluster must be modelled on a character to generate animation, as is done to produce traditional static viseme animation.

6.2 Identifying Visual Gestures

The most common approach for segmenting speech is via the acoustic modality, based on the acoustic boundaries of the uttered phones [58, 111]. However, visual gestures often overlap these boundaries as the articulators are required to be positioned prior to the onset of the sound and may remain at a position after the offset. That is, acoustic and visual speech are asynchronous. Therefore, to identify visual gestures, the speech is segmented on the visual information. This approach is sim-

ilar to that taken by [9] in which they attempt to model non-speech related facial movements in a video sequence by segmenting the motion into sub-trajectories corresponding to distinct actions. They did this by locating nodes that correspond to the areas of high density in feature space, and segmenting on these. However, these nodes marked the locations that were most frequently visited during facial motion, and for speech, they would likely correspond to the points that are frequently hit during the transition from one articulatory target to another. Instead, it is desirable to locate the boundaries at points that are visually salient, such as the peak of a mouth opening or a lip closure.

The AAM parameter trajectories are segmented into sequences of discrete, non-overlapping visual gestures, where the i^{th} gesture in a sequence, $\mathbf{G}_i$, is a sequence of feature vectors that map a trajectory in AAM space representing a distinct movement of the visible speech articulators. The boundaries between gestures are automatically defined as salient points along the trajectory, which are identified by differentiating the gradient magnitude in 20D AAM parameter space, $|\nabla_c|$, and locating the zero-crossings from negative to positive:

$$\frac{d(|\nabla_c|)}{dt} = 0, \qquad \text{where } \frac{d^2(|\nabla_c|)}{dt^2} > 0. \tag{6.1}$$

The motivation for identifying gesture boundaries in this way is that during speech the articulators do not move at a constant rate. Rather, they tend to accelerate away from articulatory targets and then decelerate as they approach the next target. Segmenting in this way generates a visually intuitive and compelling segmentation, marking boundaries where the articulators change direction, or where they hit extreme poses, such as the lip closure during a bilabial. Figure 6.1 illustrates the automatically derived visual boundaries alongside the asynchronous phone boundaries for the utterance “Would a blue feather in a man’s hat make him happy all day?”.

The gestures defined in this way are not intrinsically tied to the underlying phoneme string. Indeed the number of visual gestures in a sequence is highly unlikely

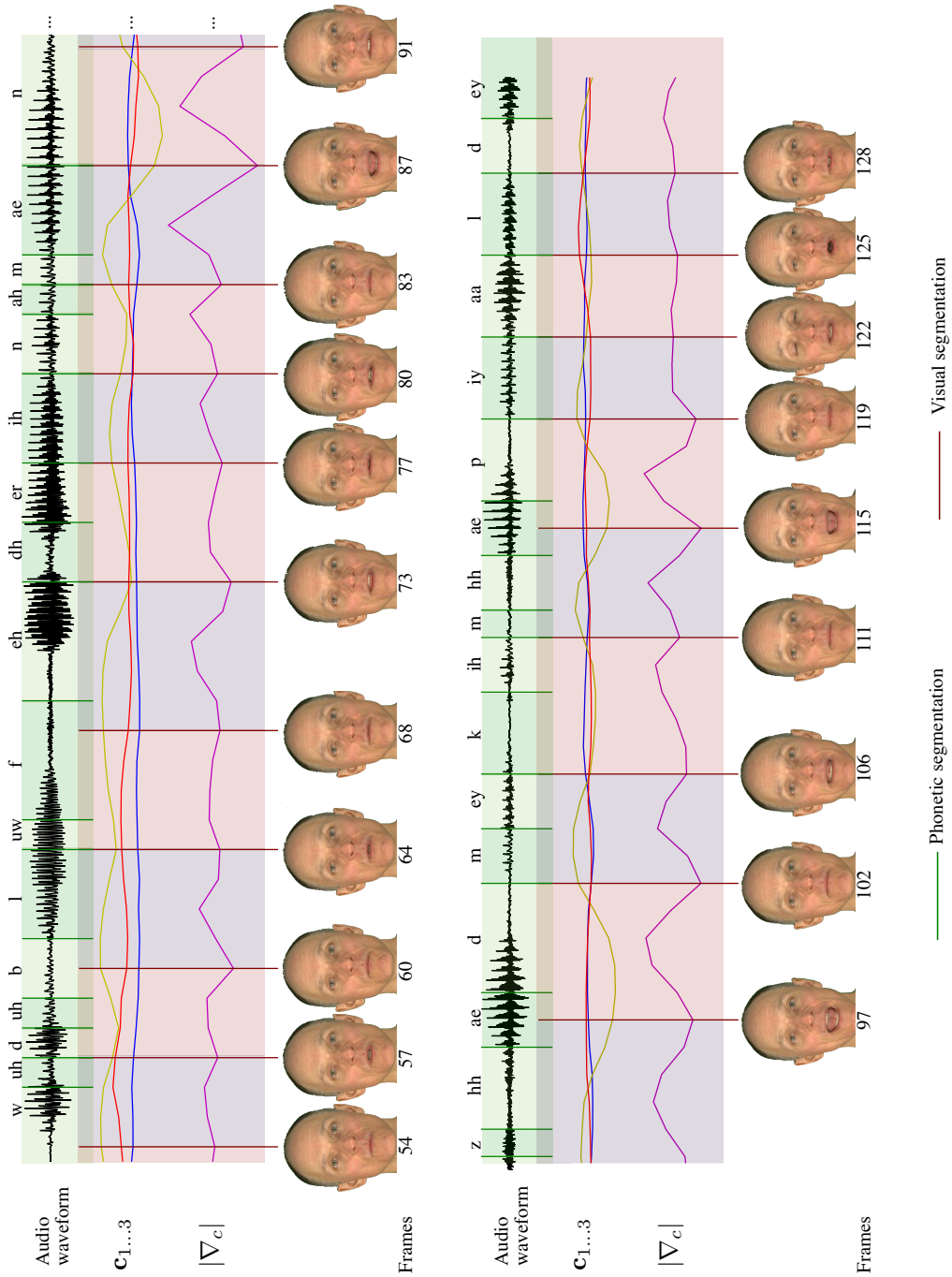


Figure 6.1: The time-varying trajectory of an audio waveform, the first three AAM parameters, and the derivative of the gradient magnitude for a sentence. The phonetic segmentation is shown in green, and the automatically derived gesture boundaries are shown in red. The video frames corresponding to the segment boundaries are displayed below the graph.

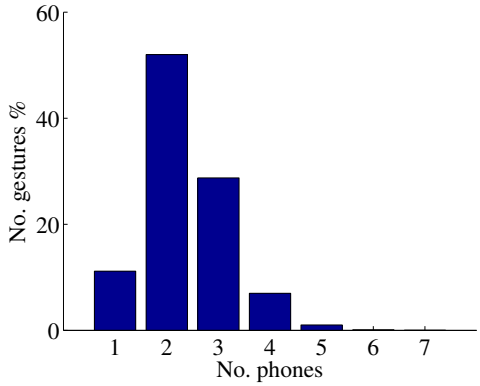


Figure 6.2: The distribution of the number of phones spanned by a visual speech gesture.

to equal the number of phones, and each sentence in the KB-2k dataset contains, on average, 35 phones and only 21 visual gestures. This is to be expected since not all sounds require an action of the visible articulators. For example, if the position of the lips has minimal influence on the articulation of a phone, then they are likely to remain in the position of the previous phone or move in anticipation to an upcoming phone — Figure 2.6 shows this. Therefore, there are generally fewer visual gestures in a sentence than phonemes.

The distribution of the number of phones per gesture calculated from the entire KB-2k dataset is presented in Figure 6.2. It shows that $\approx 90\%$ of the visual speech gestures extend over two or more phones, and $\approx 0.1\%$ span six or more phones. The frames from a six phoneme gesture can be seen in Figure 6.3 from which it is clear that although six sounds are uttered, only one distinct movement of the lips is apparent. It should be noted that using traditional animation methods, such as keyframe interpolation, seven target poses would have appeared within this sequence, which would clearly cause over-articulated animation.

The next task is to cluster the collection of variable-length, dynamic visual speech gestures from the training video into visually similar groups of gestures. Rather than referring to visemes as the visually contrastive phonemes as is traditionally done, a viseme is instead defined as the dynamic gestures that have the same visual function. The dynamic viseme groups represent meaningful contrasts between distinct visual

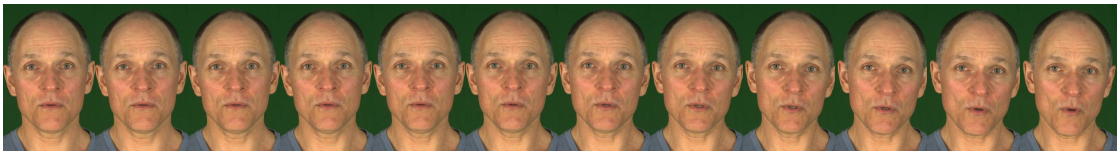


Figure 6.3: Video frames of a gesture that spans six phones and three frames of silence.

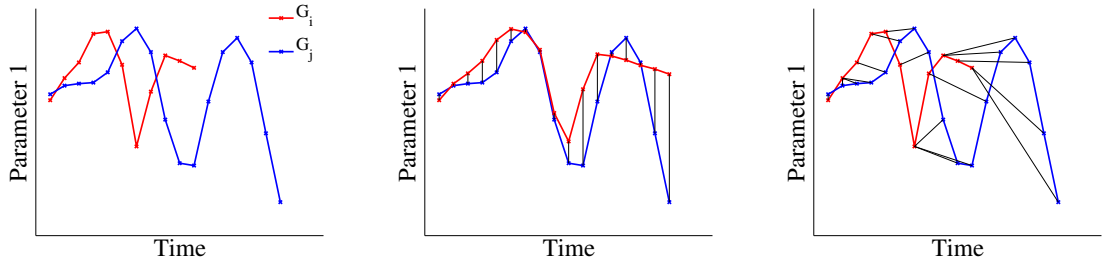
speech movements, and the gestures within a group are the visual analogue to the allophones of a phoneme.

6.3 Clustering Visual Gestures into Dynamic Visemes

The entire speech corpus can be represented as a series of T gestures, where the i^{th} gesture, $\mathbf{G}_i = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k\}$, represents a time-varying sequence of AAM parameters of duration k frames, where k naturally varies from one gesture to the next according to the segmentation described in the previous section. Dynamic viseme units are then found by clustering the gestures into visually similar groups.

To do this, initially it is necessary to determine how the distance between each pair of gestures is to be measured. It is not possible to simply calculate the frame-wise distance as the gestures are of arbitrary length, so the methods considered in this section include linearly resampling the gestures to the same length and then computing the frame-wise distance, dynamic time warping from one gesture to another, and mapping the gestures to hidden Markov model (HMM) super-feature space and computing the distance between super-features. The gestures are then clustered based on the distances identified by optimising some criteria, and the number of clusters is chosen by trading-off the number of clusters with the quality of the clusters.

Clustering is a domain dependent problem since the distribution of the data



(a) Variable length trajectories (b) Linear resampling (c) Dynamic time warping

Figure 6.4: Methods for comparing variable length time series data.

and the intended purpose of the clustering affects the perceived quality of the groups [151]. There are a large number of clustering techniques available, so it is important to select a distance function and clustering algorithm that is able to generate perceptually relevant clusters in the context of visual speech.

6.3.1 Linear Resampling

The simplest method for comparing two variable length, multivariate gestures, $\mathbf{G}_i = \{\mathbf{c}_{i,1}, \mathbf{c}_{i,2}, \ldots, \mathbf{c}_{i,k}\}$ and $\mathbf{G}_j = \{\mathbf{c}_{j,1}, \mathbf{c}_{j,2}, \ldots, \mathbf{c}_{j,l}\}$, is to linearly resample the data to a fixed length and then compute the sum of the point-wise distances. Figure 6.4(b) illustrates this process for a univariate trajectory. The distance function can be, but is not restricted to, the Euclidean distance and this will be discussed later in the chapter.

The advantages of this approach are its simplicity and efficiency. However, Figure 6.5 illustrates a limitation of this method, which is that the direction of movement is not accounted for since the distance between two trajectories travelling in opposite directions is equivalent to the distance between two trajectories which are travelling in the same direction, where one is shifted along the y axis. However, in terms of visual speech gestures, it is likely that the latter gestures are perceptually more similar to one another.

A further downfall with linear resampling is that when two trajectories are similar

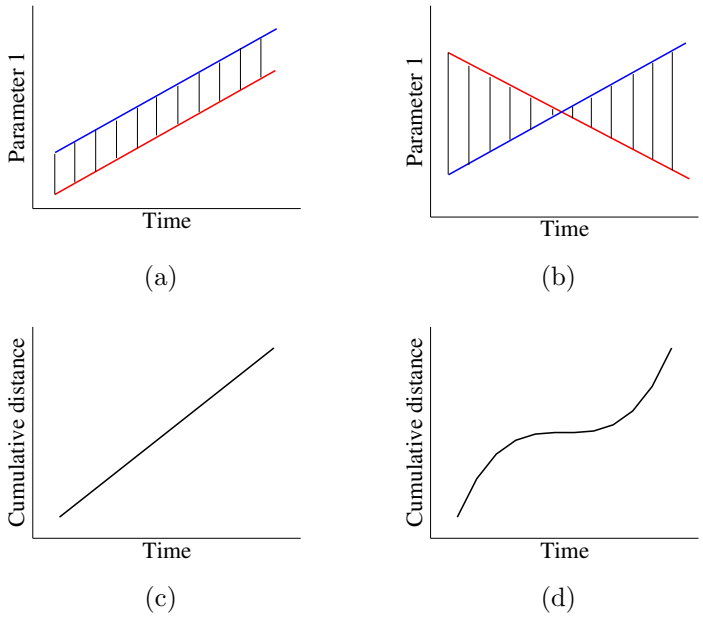


Figure 6.5: A limitation of linearly resampling gestures to uniform length, and calculating the point-wise distance. Figure (c) illustrates the cumulative distance between the trajectories shown in Figure (a) and Figure (d) illustrates the cumulative distance between the trajectories shown in Figure (b). They both produce the same distance, however, perceptually the trajectories in Figure (a) are more similar than those in (b).

in terms of shape, but are offset by a small amount in the temporal axis, the distance between them is likely to be large, and not representative of the perceptual similarity. This might arise because of variable speaking rate, so it is clear that a more flexible, non-linear approach would be beneficial.

6.3.2 Dynamic Time Warping

Dynamic time warping (DTW) is a method for measuring the similarity between two time series that may vary in length or speed. This is done by non-linearly warping along the time axis to align one of the sequences to the other such that the distance between them is minimised [82], as illustrated in Figure 6.4(c). The distance is calculated between the aligned trajectories on a frame-wise basis.

To calculate the alignment between two gestures, $\mathbf{G}_i = \{\mathbf{c}_{i,1}, \mathbf{c}_{i,2}, \ldots, \mathbf{c}_{i,k}\}$ and $\mathbf{G}_j = \{\mathbf{c}_{j,1}, \mathbf{c}_{j,2}, \ldots, \mathbf{c}_{j,l}\}$, a $k \times l$ matrix is constructed, where element (u, v) contains

the distance between the vectors $\mathbf{c}_{i,u}$ and $\mathbf{c}_{j,v}$. A warping path, $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_P\} \in \mathbb{Z}^2$, where $\max(k,l) \leq P \leq (k+l-1)$, defines the mapping between the trajectories. More specifically, the p^{th} element of $\mathbf{W}$ defines the mapping from frame $w_{p,1}$ in $\mathbf{G}_i$ to frame $w_{p,2}$ in $\mathbf{G}_j$. The warping path is calculated by traversing the distance matrix while satisfying the following conditions [82]:

- Boundary conditions: The warp path must begin at the first elements of both time-series and end at the last elements, such that $\mathbf{w}_1 = (1,1)$ and $\mathbf{w}_{\max(k,l)} = (k,l)$.
- Continuity conditions: The warp path can only move to adjacent, or diagonally adjacent cells in the matrix.
- Monotonicity conditions: The warp can only move forward or remain static in time.

There are an exponential number of paths that satisfy the above criteria. However, the *optimal* warping path is the one that minimises the overall distance between the aligned trajectories. To determine this path, a dynamic programming algorithm is used to a evaluate a cumulative distance matrix $\mathbf{\Gamma}$, where $\boldsymbol{\gamma}_{u,v}$ is the sum of the distance between $\mathbf{c}_{i,u}$ and $\mathbf{c}_{j,v}$, and the minimum of the previous cumulative distances [82]:

$$\boldsymbol{\gamma}_{u,v} = d(\mathbf{c}_{i,u}, \mathbf{c}_{j,v}) + min(\boldsymbol{\gamma}_{u-1,v-1}, \boldsymbol{\gamma}_{u-1,v}, \boldsymbol{\gamma}_{u,v-1}). \tag{6.2}$$

The overall DTW cost can be denoted:

$$\text{DTW}(\mathbf{G}_i, \mathbf{G}_j) = \frac{\boldsymbol{\gamma}_{k,l}}{P}, \tag{6.3}$$

where P normalises for path length.

The warp between the trajectories in Figure 6.5(a) can be seen in Figure 6.6. Note that the peaks and valleys of $\mathbf{G}_i$ are in alignment with those of $\mathbf{G}_j$, enabling a more intuitive comparison.

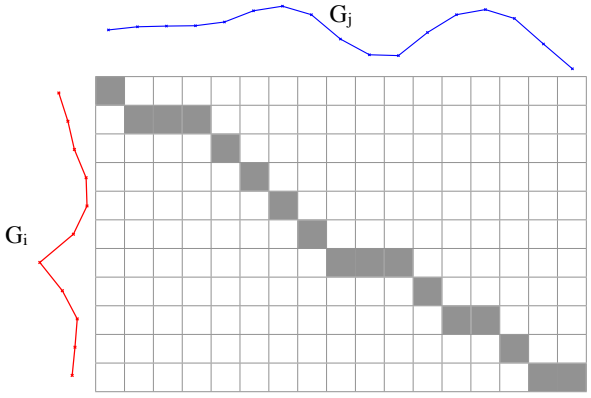


Figure 6.6: The non-linear, dynamic time warp between two trajectories.

A problem with DTW is that the run-time and space complexity is O(kl). However, as the length of the visual speech gestures in the KB-2k database is on average four frames, this remains a tractable solution.

6.3.3 HMM Super-features

An alternative way of measuring the distance between two gestures is using HMM super-features (also called super-vectors) [23], as they provide a means for representing trajectories of arbitrary duration with a fixed length feature vector. HMM super-features extend Gaussian mixture model (GMM) super-features, which are currently state-of-the-art for audio [23, 131], and audio-visual [19, 155] speaker verification. However, GMM super-features are most useful in applications where only the *manner* in which a person speaks or moves is important, and not the speech content or meaning of the gesture. This is because GMMs ignore the order of the data, and thus temporal information is lost. HMM super-features are better suited to applications for which the order is also important as temporal ordering is preserved, and they have been used for applications such as text dependent speaker verification [39].

To generate super-features, a universal background model (UBM) in the form of a hidden Markov model (HMM) is first trained using the AAM parameterisation

from all training gestures, $\mathbf{G} = \{\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_T\}$. A HMM can be defined by the following parameters [129]:

- The number of states, N
- A state transition matrix, $\mathbf{A} = \{a_{ij}\}_{ij=1}^{N}$, where $a_{ij} = P(s_j|s_i)$
- A vector of observation probabilities for each emitting state, $\mathbf{B} = \{b_j(\mathbf{G})\}_{j=1}^{N}$
- A vector of initial state probabilities, $\mathbf{\Pi} = \{\pi_i\}_{i=1}^{N}$

The parameter set can therefore be compactly written:

$$\lambda = (\mathbf{A}, \mathbf{B}, \mathbf{\Pi}). \tag{6.4}$$

The output probability distributions, $\mathbf{B}$, can be discrete or continuous depending on the observations. As the AAM parameterisation is real-valued, the models used in this work are continuous density HMMs (CD-HMMs), where each state, $b_j(\mathbf{G})$, is represented as a multivariate Gaussian mixture model (GMM):

$$b_j(\mathbf{G}) = \sum_{m=1}^{M} \phi_{jm} \mathcal{N}(\mathbf{G}; \boldsymbol{\mu}_{jm}, \Sigma_{jm}), \tag{6.5}$$

where $\mathcal{N}(\mathbf{G}; \boldsymbol{\mu}_{jm}, \Sigma_{jm})$ denotes a multivariate Gaussian with mean $\boldsymbol{\mu}_{jm}$ and covariance Σ_{jm}, M is the number of mixture components and ϕ_{jm} is the weight of the m^{th} mixture component.

The UBM HMM is trained iteratively using the Expectation-Maximisation (EM) algorithm [37] with the maximum-likelihood criterion, where the goal is to compute:

$$\Lambda^* = \arg\max_{\Lambda} P(\mathbf{G}|\Lambda), \tag{6.6}$$

where

$$P(\mathbf{G}|\Lambda) = \prod_{t=1}^{T} P(\mathbf{G}_t|\Lambda). \tag{6.7}$$

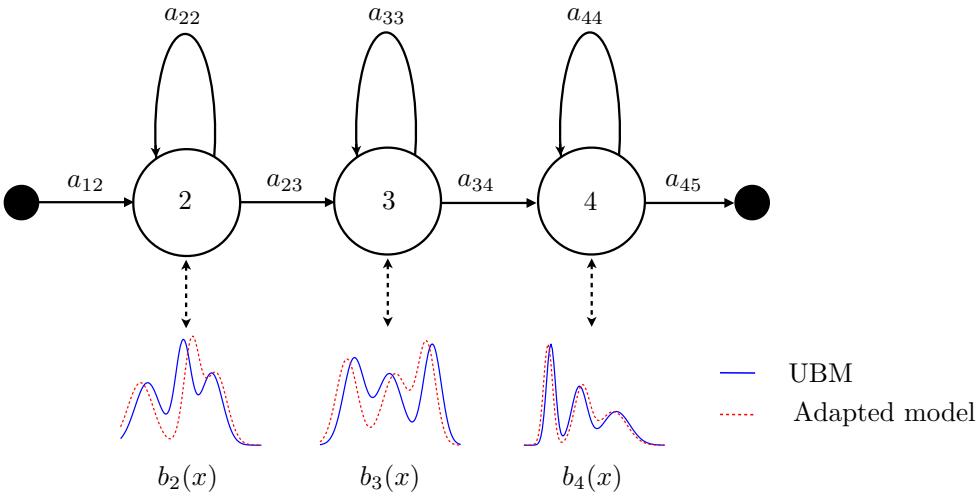


Figure 6.7: A UBM trained using every speech gesture in the training data in the form of a five state (three emitting state) left-to-right HMM where each state is modelled with a multivariate GMM. The means of each mixture component are then individually adapted for each speech gesture. The first GMM for each state of the UBM is shown in blue and the adapted model is shown in red.

This generates a gesture-independent representation of the speech movements. For each gesture, the UBM is then adapted using maximum a posteriori (MAP) estimation [159], and the means of the mixture components are updated using:

$$\hat{\boldsymbol{\mu}}_{jm} = \frac{N_{jm}}{N_{jm}+\tau}\bar{\boldsymbol{\mu}}_{jm} + \frac{\tau}{N_{jm}+\tau}\boldsymbol{\mu}_{jm}, \tag{6.8}$$

where $\boldsymbol{\mu}_{jm}$ is the mean of the j^{th} state and m^{th} mixture component of the UBM, $\bar{\boldsymbol{\mu}}_{jm}$ is the mean of the adaptation data, τ is the weight of a priori knowledge to the adaptation data, N_{jm} is the occupation likelihood of the adaptation data and $\hat{\boldsymbol{\mu}}_{jm}$ is the updated mean. An example of a 5 state HMM with 3 Gaussian mixture components per state can be seen in Figure 6.7, where the blue curves illustrate the means and variances of the mixture components for the first GMM of the UBM and the red curve shows the means after adaptation.

The HMM super-features for gesture i, $\mathbf{g}_i^s$, are the stacked vector difference be-

tween the UBM mean vectors and the MAP adapted mean vectors:

$$\mathbf{g}_i^s = \begin{pmatrix} \boldsymbol{\mu}_{1,1} \\ \boldsymbol{\mu}_{1,2} \\ \boldsymbol{\mu}_{1,3} \\ \vdots \\ \boldsymbol{\mu}_{N,M} \end{pmatrix} - \begin{pmatrix} \hat{\boldsymbol{\mu}}_{1,1} \\ \hat{\boldsymbol{\mu}}_{1,2} \\ \hat{\boldsymbol{\mu}}_{1,3} \\ \vdots \\ \hat{\boldsymbol{\mu}}_{N,M} \end{pmatrix} \tag{6.9}$$

The dimensionality of the super-features is $N \times M \times D$, where N is the number of states, M is the number of Gaussian mixture components and D is the dimension of the AAM parameterisation. The models described in this work are trained using the AAM parameters appended with the velocity (Δ) and acceleration ($\Delta\Delta$) coefficients making $D = 20 \times 3 = 60$. The values N and M are defined in Section 6.3.5.

The UBM HMM is a left-to-right model with self-looping allowed, but no state skipping as the parameters of visual speech change in a successive manner. The HMMs are trained and adapted using the algorithms from the Hidden Markov Model Toolkit (HTK) [159].

6.3.4 Selecting a Distance Function

There are many methods of calculating the proximity between sets of features, and different measures can lead to very different solutions. For this work it is important that the distance between visual speech gestures is indicative of the perceptual similarities between the gestures. Those gestures that look very similar should have a low distance and vice-versa. To determine the best distance function for the KB-2k dataset, the efficacy of a variety of measures was determined by comparing against a set of subjective judgements.

6.3.4.1 Subjective Distances

To gather perceptual judgements, four participants were asked to observe a series of gestures and select the gestures that they perceived to be the same movement of the speech articulators. Due to the large number of comparisons necessary for pair-wise judgements across all of the gestures in the dataset, a subset of 260 were randomly selected. Of the 260, 10 were selected as *reference* gestures, against which the remaining 250 *test* gestures were compared. A graphical user interface displayed 5 of the 10 reference gestures along the top of the screen, and 1 of the 250 test gestures below. The user was automatically prompted to select which of the 5 reference gestures they deemed to match the test gesture in terms of visual meaning. An option was offered for no match and multiple matches were allowed. Once rated, the participant could proceed to the next gesture, and the process was repeated until all of the 250 gestures had been compared against each of the 10 reference gestures.

During this process, the gesture movies were presented to the participants with no audio, and all movies on the screen were played simultaneously. Participants were able to play the gestures as many times as necessary. Due to the large number of comparisons, all four participants completed the task over multiple sessions, each lasting on average 20–25 minutes.

The participants' scores were collated and the number of times each test video was marked as similar to each reference video was counted and normalised to the range $[0, 1]$. Figure 6.8 shows a visualisation of the distribution of the similarity judgements over all participants. For reference gesture $\mathbf{G}_i$, the normalised scores can be represented with the vector $\mathbf{q}_i = \{q_{i,1}, q_{i,2}, \ldots, q_{i,250}\}$, where $q_{i,j} \in \{0, 0.25, 0.5, 0.75, 1\}$.

6.3.4.2 Comparing Objective and Perceptual Distances

The subjective judgements are used to measure the quality of the various objective distance measures that can be used for clustering. For each of 12 distance metrics

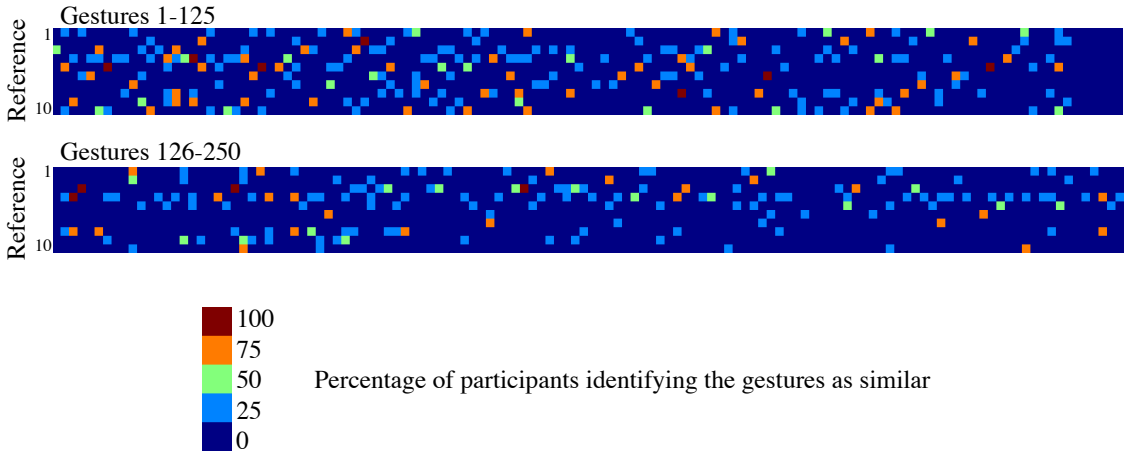


Figure 6.8: Visualising the similarity judgements across four participants on comparing ten reference gestures (rows) to 250 other randomly selected test gestures (columns).

listed in Appendix B, the distance between the 10 reference gestures and 250 test gestures is calculated. To account for the variability in gesture length, the three methods described previously are used: (a) Linear resampling to the longest gesture (LR), (b) Dynamic time warping (DTW), and (c) Mapping the gestures to fixed-length HMM super-features (SF). In each case, the velocity and acceleration coefficients are appended to the AAM parameters.

A problem with directly comparing the objective and subjective results is that they represent different types of data. The collated subjective scores are naturally quantised, whereas the objective scores are real numbers corresponding to the distance between pairs of gestures in feature space. To accommodate this, one approach is to map the distances to a smaller set of values with the same number of quantisation levels as the subjective data. During this process it is also important that the proportion of the gestures at each quantisation level for the subjective data matches that for the quantised objective data. For example, if 10% of the scores were equal to quantisation level one in the subjective data, the same proportion of the objective data should be quantised to level one. Spearman's rank correlation could then be used to measure the correlation between the two vectors. However, as the majority of the values are zero since each test gesture will typically look like

only one reference gesture, this method returns a very high correlation coefficient, which is not reflective of the quality of the distance metric.

An alternative approach is based on information retrieval. First, for each reference gesture, $\mathbf{G}_i$, the subjective scores are converted to a binary vector, $\mathbf{qb}_i$, such that

$$qb_{i,j} = \begin{cases} 0 & \text{if } q_{i,j} < 0.75 \\ 1 & \text{if } q_{i,j} \geq 0.75 \end{cases}. \tag{6.10}$$

This is based on a majority-vote scheme, whereby a test gesture is marked as equivalent to a reference gesture if 3 or more of the 4 participants agree that the gestures appear to portray the same visual meaning. Based on this assumption, each reference gesture has, on average, 9 test gestures that consist of the same action on the lips. The index positions of the non-zero elements of $\mathbf{qb}_i$ are stored in $\mathbf{qb}_i^{\text{ind}}$ such that

$$\mathbf{qb}_i^{\text{ind}} = \{j : q_{ij} = 1, \forall q_{i,j} \in \mathbf{q}_i\} \tag{6.11}$$

For a reference gesture, $\mathbf{G}_i^r$, the distance to the test gestures, $\mathbf{G}_{1...250}^{nr}$, is represented with the vector, $\mathbf{v}_i = d(\mathbf{G}_i^r, \mathbf{G}_{1...250}^{nr}) = \{v_{i,1}, v_{i,2}, \ldots, v_{i,250}\}$ which is then sorted into ascending order, $\mathbf{v}_i' = sort(\mathbf{v}_i)$. The index positions of the elements from $\mathbf{v}_i$ in the sorted list $\mathbf{v}_i'$ are represented by $\mathbf{v}_i^{\text{ind}}$, such that

$$\mathbf{v}_i' = \mathbf{v}_i[\mathbf{v}_i^{\text{ind}}] \tag{6.12}$$

A distance function that perfectly reflects the perceptual judgements would contain the values of $\mathbf{qb}_i^{\text{ind}}$ in the first elements of $\mathbf{v}_i^{\text{ind}}$. The similarity of the objective and subjective scores can therefore be calculated by the precision, $\mathbf{p}$, and recall (or specificity), $\mathbf{r}$:

$$p_{i,t} = \frac{|\mathbf{qb}_i^{\text{ind}} \cap \mathbf{v}_{i,1...t}^{\text{ind}}|}{t} \tag{6.13}$$

and

$$r_{i,t} = \frac{|\mathbf{qb}_i^{\text{ind}} \cap \mathbf{v}_{i,1...t}^{\text{ind}}|}{|\mathbf{qb}_i^{\text{ind}}|} \tag{6.14}$$

where t is the rank in the sorted vector. Precision calculates the proportion of gestures in rank t or above that are similar to gesture i . Recall is the proportion of all gestures that are similar to gesture i that are in rank t or above. To express the performance of the distance functions as a single number, the F-score (or F-measure) statistic is used [66]. The F-score is the harmonic mean of the precision and recall and can be calculated as follows:

$$f_{i,t} = 2\frac{p_{i,t}r_{i,t}}{p_{i,t} + r_{i,t}} \tag{6.15}$$

The F-score is defined for all 250 points along the precision and recall vectors. A standard approach is to report the maximum F-score, which is a value in the range [0,1], where 1 corresponds to a ranked vector that unequivocally agrees with the perceptual ranking. The maximum F-score is calculated for each distance function and averaged over the reference gestures. The results are shown in Figure 6.9. The results show that, although the benefit is not significant, measuring the Euclidean distance in super-feature space provides a proximity function that most closely matches the perceptual judgements.

6.3.5 Clustering Algorithms

Cluster analysis is used for discovering groups of observations in data, such that the observations within a class are more similar than those across different classes. For this work, the gestures with the same visual meaning are clustered to form a reduced set of visual speech actions which are referred to as dynamic visemes. The idea is that the gestures within a dynamic viseme group all appear to be producing the speech movements with the same visual function.

There are a large number of established clustering algorithms, including partitional, agglomerative and graph-based methods. The most well-known example of a partitional clustering algorithm is k-means. k-means and other partitional algorithms begin by modelling all observations as a single cluster. At each iteration

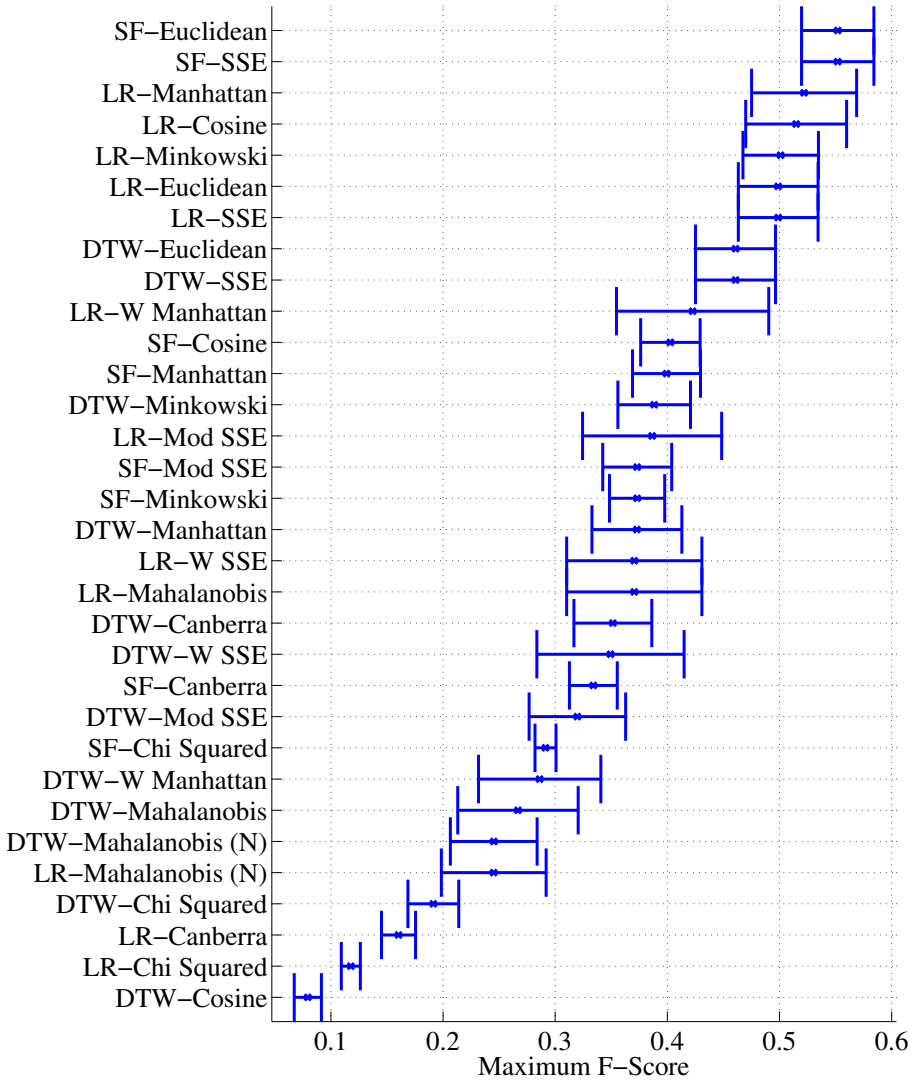


Figure 6.9: The mean and standard error of the maximum F-score averaged over all reference gestures for the different distance measures. The DTW prefix corresponds to dynamic time warping, LR to linearly resampling and SF to super-features.

the clusters are partitioned to maximise (or minimise) some clustering criterion. Conversely, agglomerative methods begin by assigning each observation to its own cluster, and then iteratively join clusters until convergence. For both of these methods, the cluster criteria often involve maximising the intra-class similarity and/ or minimising the inter-class similarity, and the algorithm converges when either some cluster quality threshold has been satisfied, or the number of clusters, k, specified by the user, has been reached.

To perform graph-based clustering, each observation is represented as a node in an undirected similarity graph. Edges connect each node to its n-nearest neighbours. The graph is then partitioned to minimise some optimisation criteria, such as minimising the intersected edges, or maximising the distances assigned to each edge. This graph is recursively partitioned until the user-defined number of clusters has been reached.

There is no generic clustering algorithm that performs best under all conditions. That is, for various types of data and applications, different clustering algorithms are more suited. Therefore, a variety of algorithms were used to cluster the previously identified visual speech gestures using the CLUstering TOolkit (CLUTO) [77]. CLUTO is stand-alone software which contains several implementations of partitional, agglomerative and graph-based clustering algorithms.

The Euclidean distance between gestures in super-feature space is used as the basis for the clustering. The UBM that is used for deriving the super-features is a five state (three emitting state) HMM, with a single mixture component for each state. This topology was chosen as it attained the maximum F-score in the previous section. This is likely to be because there are only a small number of frames per gesture. Those gestures that are shorter than three frames are omitted from analysis. However, these account for less than 5% of the data.

After clustering using each algorithm, the movie frames corresponding to the clustered gestures were visually inspected and it was established that the graph-based clustering algorithm generates visually better clusters than the other methods. A

possible reason for this is that graph-based algorithms are naturally able to generate non-spherical clusters, so can more accurately model data that does not conform to a normal distribution.

The graph is formed by connecting each node (gesture) to its 40 nearest neighbours and is partitioned using the multi-level, recursive bisections algorithm described in [78]. First the graph is coarsened to a few hundred nodes, then the coarse graph is bisected while maximising the edge weight, and finally, the bisection is projected onto the original, finer graph, while refining the partition. This produces a good quality clustering that is an order of a magnitude faster than traditional recursive bisection algorithms. All of the sentences from the KB-2k dataset were clustered with the exception of a random set of 50 sentences which were held out for testing (see Section 7.3).

6.3.6 How Many Clusters?

The algorithm used for clustering requires that the number of clusters is given a priori, and converges when that number of clusters is reached. There is an abundance of methods to determine the number of clusters, k, that should be used to capture the distribution of a set of data. Typically, they involve calculating a measure of cluster quality for varying k, and plotting this value against k. This plot describes the trade off between the number of groups and the quality of the clusters, so where a large change in the gradient of the curve occurs, the value of k is suggestive of the number of clusters [44, 74]. Three common cluster quality measures are:

1. Dunn's Index (DI) [61]

$$DI = \frac{d_{\min}}{d_{\max}} \tag{6.16}$$

where, $d_{\min}$ denotes the smallest distance between two gestures from different clusters, and $d_{\max}$ is the largest distance between two gestures from the same cluster. A large value indicates good clusters, as the distance between gestures from different clusters should be high, and the distance between gestures from

within a cluster should be low.

2. Davies Bouldin Index (DBI) [61]

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{j=1\ldots k, j\neq i} \frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \tag{6.17}$$

where σ_i is the average distance of the gestures in cluster i to the cluster centroid and $d(c_i, c_j)$ is the distance between the cluster centroids for clusters i and j. This measure calculates the maximum ratio between the cluster compactness and separation, so a small DBI indicates good quality clusters.

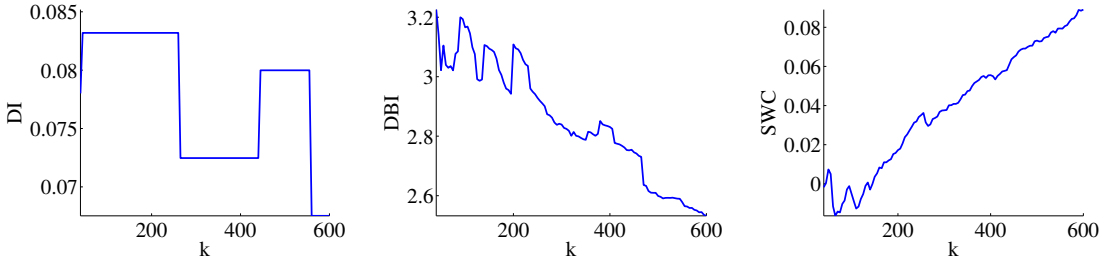
3. Silhouette Width Criterion (SWC) [135]

$$SWC = \frac{1}{t}\sum_{i=1}^{t} \frac{b_i - a_i}{\max(a_i, b_i)} \tag{6.18}$$

where a_i is the average distance between gesture i and all other gestures from the same cluster, and b_i is the lowest average distance between gestures i and all gestures from a different cluster. The SWC is a value in the range [-1, 1], where a value of 1 indicates good clusters, which occurs when a_i is smaller than b_i.

To determine the number of clusters required, and hence the number of dynamic viseme classes, the three cluster quality measures are computed for each of $k = \{40, 45, 50, \ldots, 600\}$. For each method, the score is plotted against the number of clusters, and is shown in Figure 6.10. However, from these graphs it is unclear what is the required number of clusters to model the data best. The maximum Dunn's Index value appears between $k = 5$ and $k = 255$, whereas the minimum Davies Bouldin Index, and maximum Silhouette Width Criterion appears to be at $k \geq 600$. In no case does a *knee* appear to suggest a good trade off between the number of clusters and the quality.

For a clearer measure of the cluster quality which is more suited to our application, two goodness-of-fit measures are computed for each of $k = \{40, 45, 50, \ldots, 600\}$.



(a) Dunn's Index (b) Davies Bouldin Index (c) Silhouette Width Criterion

Figure 6.10: Three common cluster quality measures plotted for each of $k = \{40, 45, 50, \ldots, 600\}$.

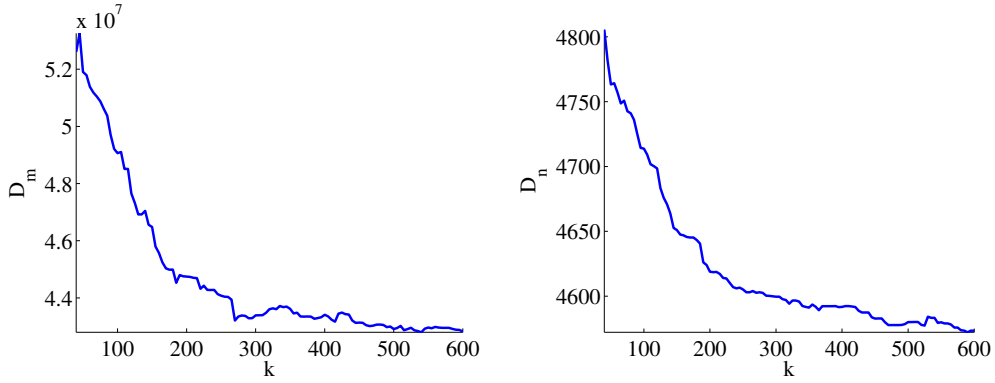
These are as follows:

D_m the mean distance of the super-features to their respective cluster median.

D_n the mean distance of the super-features to the nearest sample that does not belong to the same cluster.

Given that a cluster of gestures (a dynamic viseme class) should represent visually similar units, D_m is expected to be small if the data are clustered well as the gestures should all be close to their respective cluster centroid in super-feature space. The second measure, D_n, is included to determine if there are either too many clusters, as neighbouring clusters will contain very similar gestures resulting in a low error, or too few clusters, as the resulting error will be high as the nearest samples not in the same cluster are increasingly further away. These measures are each plotted in Figure 6.11. To determine the number of clusters required, the knee of the curves is located. In this case the knee is around 150, which is used to define the number of units.

Example clusters are shown in Figures 6.12, 6.13 and 6.14. In each case, the movie frames corresponding to a different gesture that appears within the cluster are shown on each row. Figure 6.15 shows the trajectories of the first AAM parameter corresponding to the median and the fifteen gestures that are closest to the median for each of the dynamic visemes in Figures 6.12, 6.13 and 6.14.



(a) D_m (b) D_n

Figure 6.11: The mean squared difference between the super-features and the respective cluster median for each gesture (D_m) and the nearest-neighbour from a different cluster (D_n). The number of clusters is varied over $k = \{40, 45, 50, \ldots, 600\}$. The trade-off value for k is around 150 clusters.

6.4 The Relationship Between Dynamic Visemes and Phonemes

The dynamic visemes identified by clustering represent sets of visually similar gestures. Therefore, if mapping phonemes to static visemes is valid, phonemes would be assigned to the same visual clusters consistently. However, this is far from the case. Instead, a particular phone or sequence of phones tends to appear in many different dynamic viseme clusters because they can appear very different on the lips. In Figures 6.12, 6.13 and 6.14 it is apparent that a cluster comprises of gestures produced by various phone sequences. To further illustrate this, Figure 6.16 shows the thirty most frequent phone sequences that appear within eight of the clusters. It is clear that, although some of the common groupings from the traditional phoneme-to-viseme mappings (see Figures 3.4 and 3.5) do appear in the clusters, the relationship is far more complex since, in these examples the phone sequence /ðʊ/ appears in both clusters 2 and 4, and the phone /s/ appears in isolation in clusters 1, 2, 4, 22 and 24.

Indeed, all of the phones in the training data are distributed over a large number of clusters. For example, Figure 6.17 shows that occurrences of the phonemes /ʃ/,

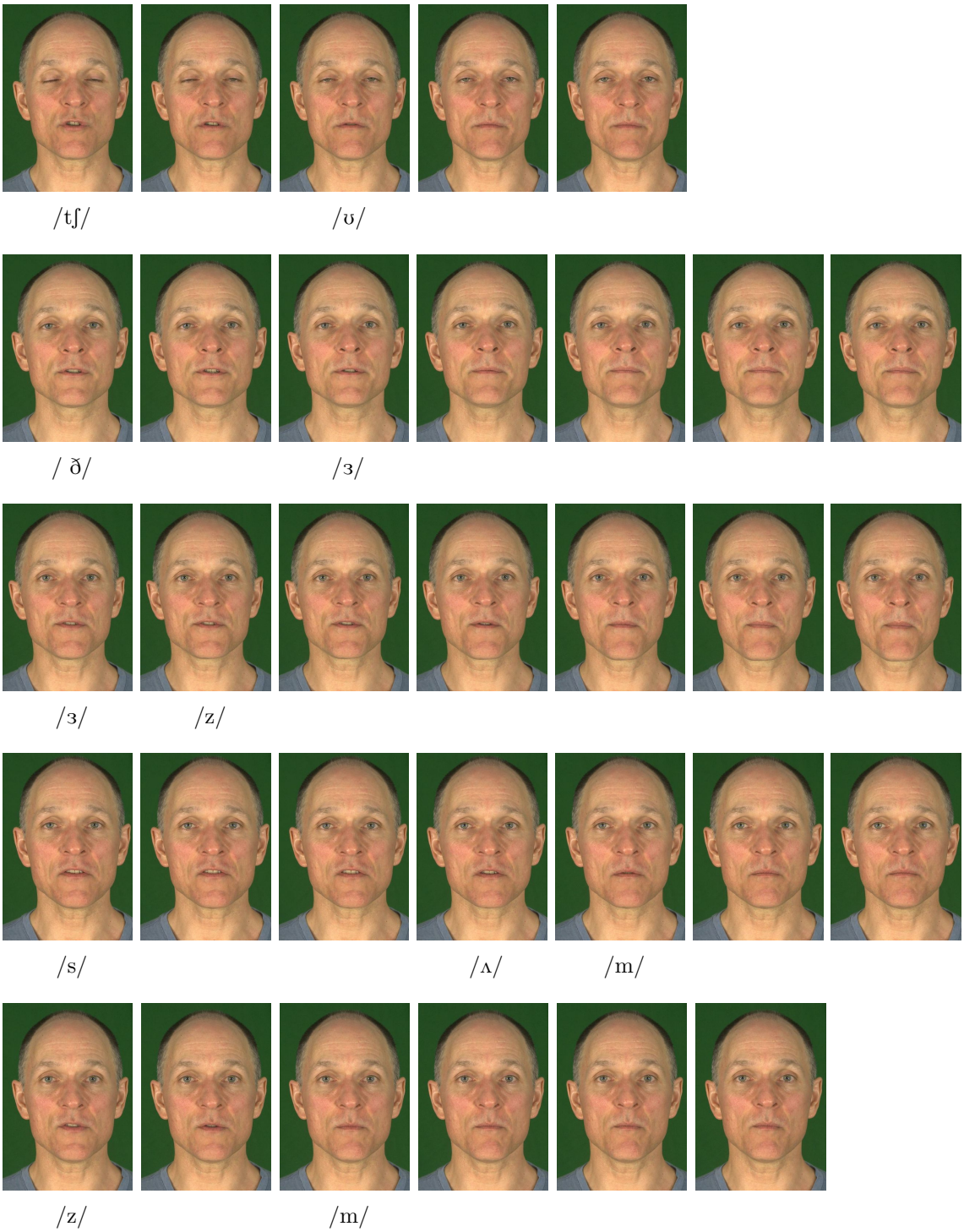


Figure 6.12: Video frames corresponding to five gestures from cluster one. Each row represents a different gesture from the cluster.

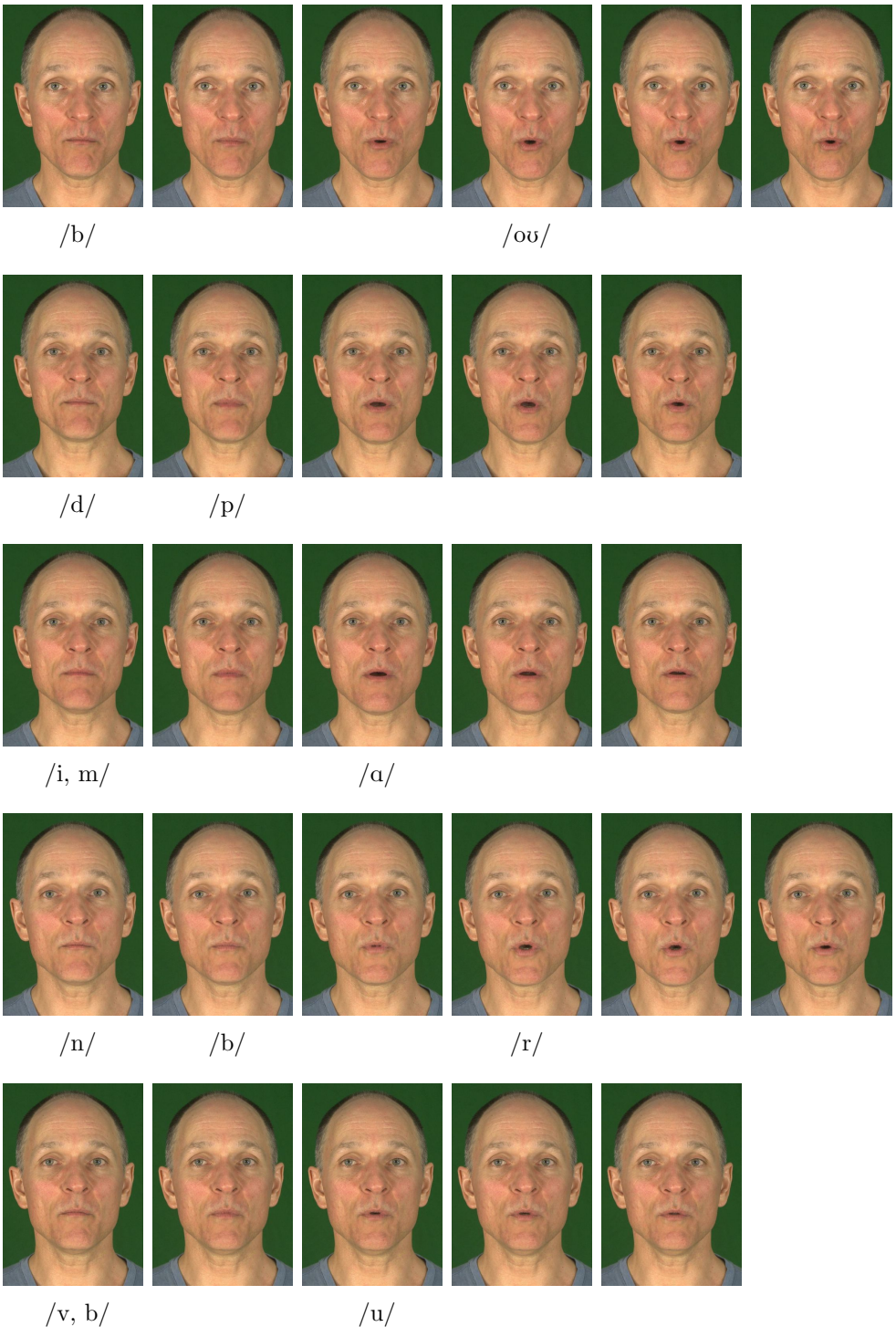


Figure 6.13: Video frames corresponding to five gestures from cluster three. Each row represents a different gesture from the cluster.

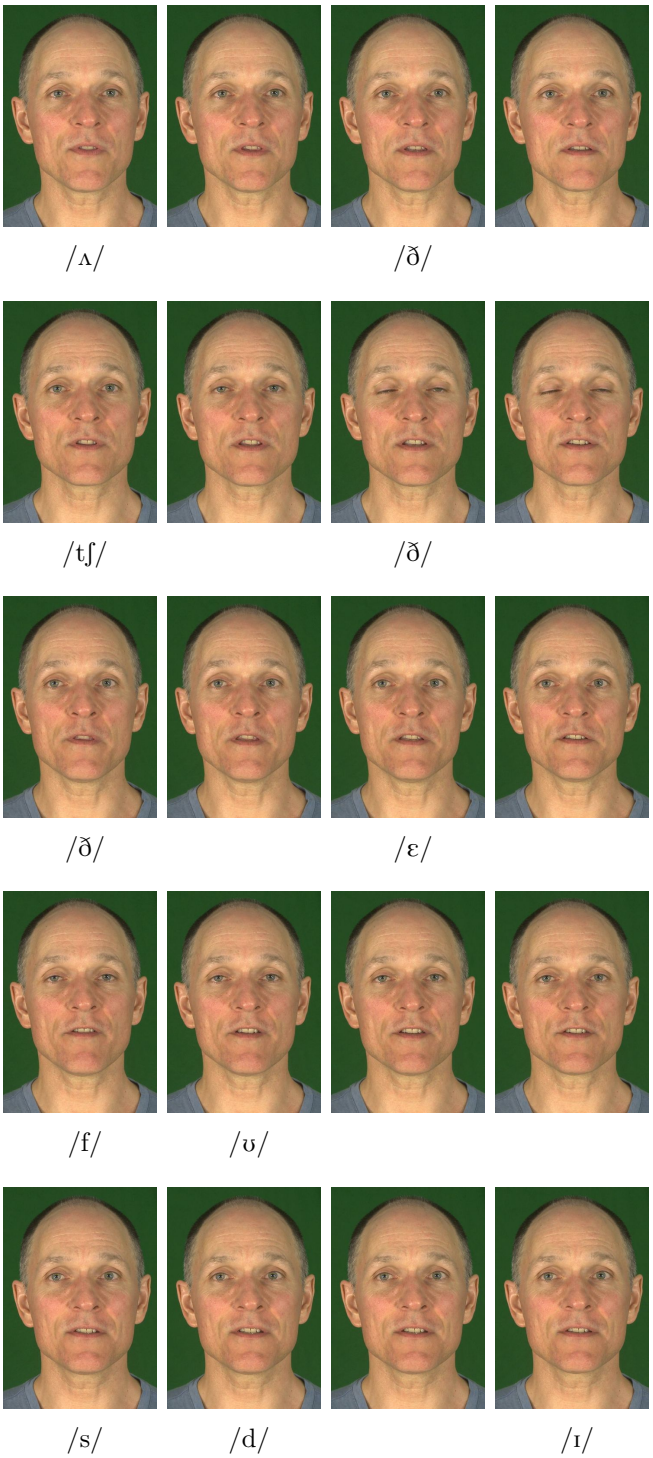
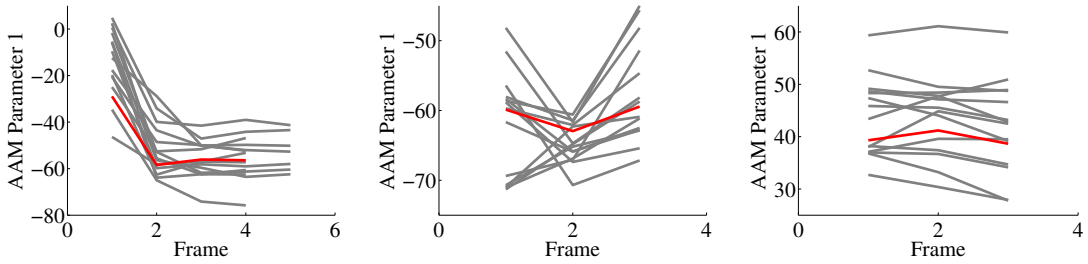


Figure 6.14: Video frames corresponding to five gestures from cluster four. Each row represents a different gesture from the cluster.



(a) Viseme 1 (b) Viseme 3 (c) Viseme 4

Figure 6.15: The trajectories of the first AAM parameter corresponding to the median and the fifteen gestures that are closest to the median for each of the visemes in Figures 6.12, 6.13 and 6.14.

and to a greater extent, /d/ are distributed widely over the 150 dynamic visemes because their visual appearance varies in different phonetic contexts. It is likely that /d/ is distributed over a wider range of clusters than /ʃ/ because it is produced by touching the tongue tip or blade against the alveolar ridge, granting freedom to the position of the lips, whereas /ʃ/ is typically produced with a degree of lip rounding. /d/ is therefore more visually deformable, so is more influenced by coarticulation. The amount of dispersion throughout the clusters for a particular phone can be measured by calculating the Shannon entropy:

$$H(p) = -\sum_{i=1}^{k} P(C_i) \log P(C_i) \tag{6.19}$$

where $P(C_i)$ is the probability that the phoneme label p appears in cluster i, and $P(C_i) \log P(C_i)$ is defined as 0 if $P(C_i) = 0$. $P(C_i)$ is calculated:

$$P(C_i) = \frac{N(p = C_i)}{N(p)} \tag{6.20}$$

where $N(p = C_i)$ is the number of times the phoneme label appears in cluster i and $N(p)$ is the total number of appearances of the phoneme. The cluster entropy for the phonemes in the KB-2k dataset is presented in Figure 6.18. Figure 6.18(a) shows that the more rounded consonants, such as /ʒ/, /tʃ/ and /w/ are least distributed throughout the clusters, closely followed by the unrounded, frontal consonants /b/,

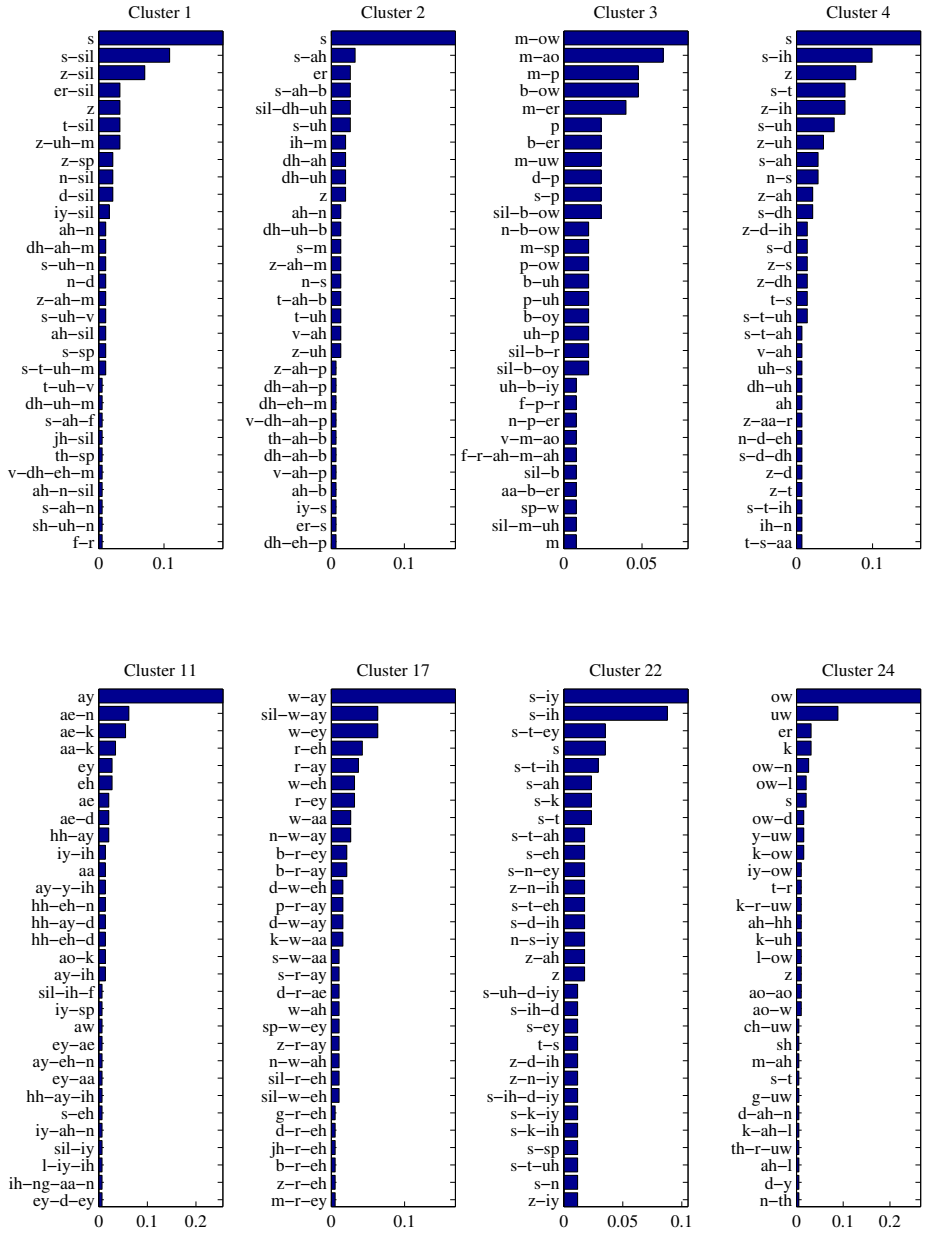
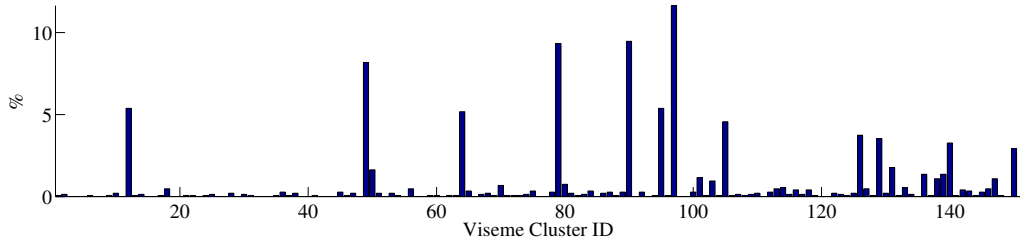
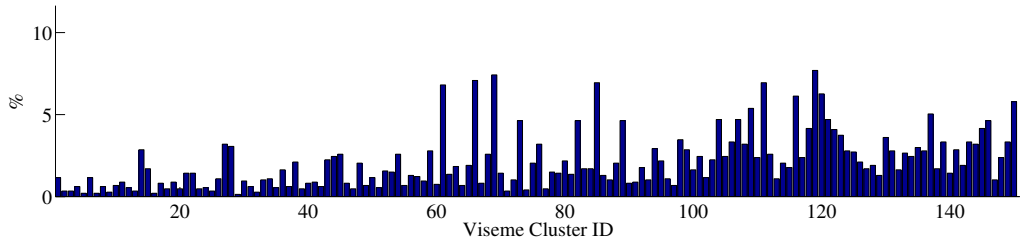


Figure 6.16: Histograms showing the twenty most frequent phoneme sequences corresponding to the clustered gestures for four dynamic visemes. In these graphs, *sil* refers to silence that occurs at the beginning or end of an utterance, and *sp* refers to a short pause that happens mid-sentence.



(a) ʃ



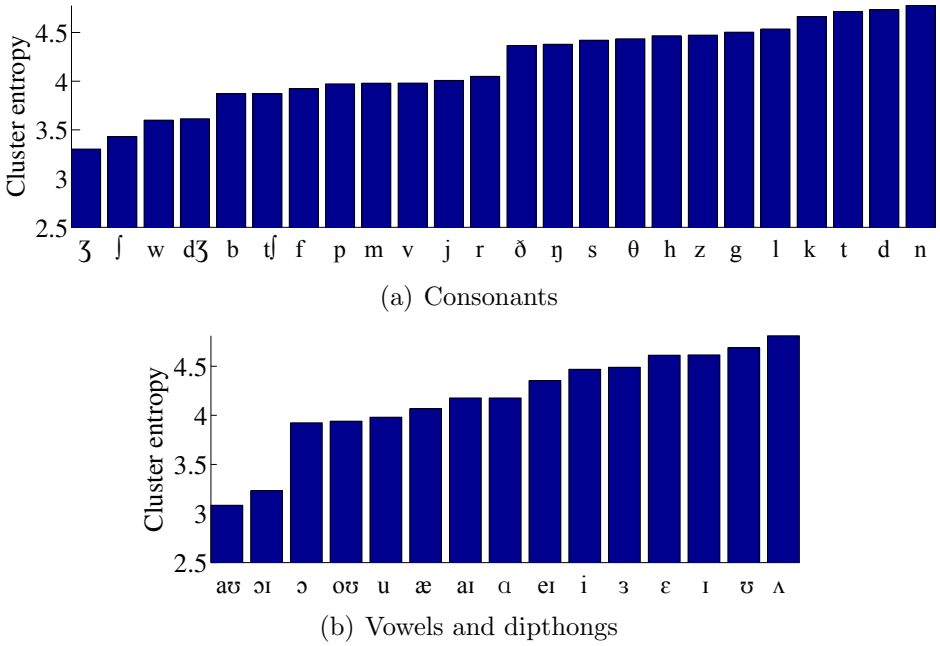
(b) d

Figure 6.17: (a) The distribution of the phoneme /ʃ/ throughout the viseme clusters. (b) The cluster distribution for the phoneme /d/. These graphs show that a many-to-one mapping from the phonemes to the visemes is not correct.

/p/, /m/ and /v/. The consonants that are most evenly distributed throughout the clusters are the glottal phones such as /g/ and /k/, and the alveolars /l/, /t/, /d/ and /n/. Alveolars involve movement of the tongue tip or blade against the alveolar ridge, which is a relatively frontal action. However, during this process, the position of the lips has little effect on the sound that is generated and so they are highly influenced by visual coarticulation.

The cluster entropy of the vowels is presented in Figure 6.18(b), and shows similarities with the consonants since rounded phones such as /aʊ/, /oʊ/ and /u/ generally have lower entropy than unrounded phones.

The findings in this section demonstrate the complex many-to-many mapping between audio and visual speech. Thus, it is unsurprising that a particular phoneme string can map to a variety of different sequences of dynamic visemes depending on the context in which phonemes appear, which is not the case for static visemes. As an example, instances of the word “another” from the KB-2k dataset are shown in Table 6.1 with their dynamic viseme transcriptions which are determined by segmenting and clustering as described in this chapter. The centre column shows



(a) Consonants

(b) Vowels and dipthongs

Figure 6.18: The entropy of the phoneme distribution throughout the dynamic viseme clusters.

the viseme sequence and the left and right columns show the context in which the word was spoken. Notice that the transcription of the word differs both in the number and in the composition of the dynamic visemes and that certain dynamic visemes tend to appear more than others. For example, when the word is preceded with silence, dynamic viseme 145 is often the first gesture that is produced, and in several different contexts, viseme 80 appears mid-way through the word where it is less likely to be affected by the neighbouring speech.

6.5 Evaluation

To evaluate the structure of the AAM parameters when grouped by phoneme, static viseme, and dynamic viseme classes, a set of ordered distance matrices were produced. Figure 6.19 presents a visualisation of the distance matrix produced by finding the distance to each pair of phonemes, and Figure 6.20 shows the distances between each pair of visemes as defined by Parke and Waters [121]. To produce this

Left context	Visemes (/ʌ-n-ʌ-ð-ɜ/)	Right context
After	70-80-124	long pause...
(Silence)	134-80-101	memo for...
... one or	83-80-149	of the...
(Silence)	134-117-35	field had...
... can have	28-80-104	tunafish sandwich
(Silence)	145-45-145-148	longer strip...
(Silence)	145-80-69	brand of...
... pick up	123-80-5	pack on ...
(Silence)	145-80-1-137	put sex...
(Silence)	145-45-67-132	snarled close...
... ideas surfeit	117-80-133	sector of...
... progress,	145-45-80-134	is delineating...
... not try	75-80-134	club
(Silence)	145-45-67-125	stock vaudeville...

Table 6.1: The centre column shows the viseme sequences for the word "another" spoken in different contexts.

image, the AAM features were segmented via the acoustic phone boundaries. Labels were then assigned to each segment based on the phoneme or the phoneme to viseme mapping taken from [121]. In these images, the samples are ordered by phoneme label, and viseme group and the class boundaries are outlined with a black box. For instance, in Figure 6.21 all segments with phoneme labels /p/, /b/ and /m/ are arranged sequentially in the distance matrix. The colours represent the Euclidean distance between the mid-frame of the phonemes measured in AAM space, ranging from blue to red representing, respectively, the smallest to largest values. A perfect clustering would therefore produce a series of blue boxes down the leading diagonal on a red background, as the distances within a cluster should be small, and across other clusters should be large.

For comparison, Figure 6.21 shows the distance matrix produced by calculating the Euclidean distance between the mid-frame of each pair of visual speech gestures for the first 40 dynamic viseme clusters[1], and ordering the samples by their cluster

[1]Only the first 40 clusters were displayed to make the visualisation clearer, and a comparison between distance matrices easier. All 3 distance matrices were subsampled to contain ≈1600 rows and columns.

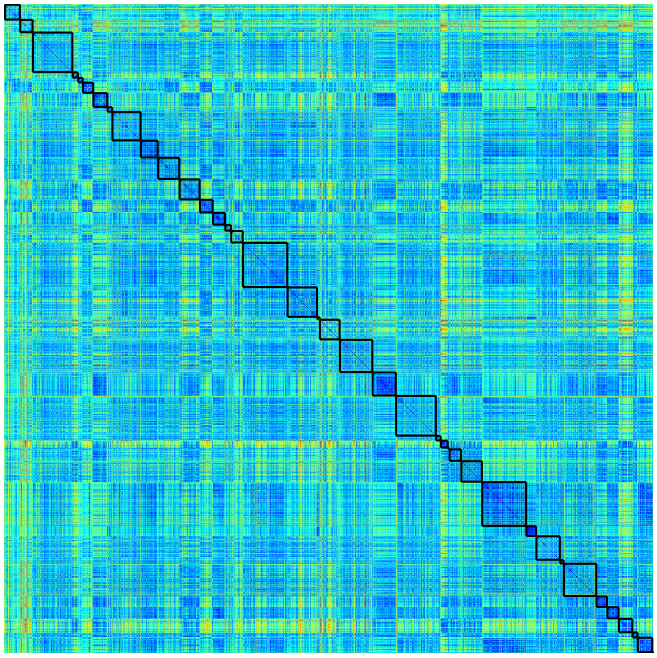


Figure 6.19: Distance matrix showing the Euclidean distance between the combined AAM parameters at the mid-frame of each phone, ordered by phoneme label. The phoneme groups are highlighted with the black boxes along the leading diagonal.

ID. The cluster boundaries are highlighted with black boxes along the leading diagonal. It is clear that the distance matrices that are based on the acoustically segmented visual features appear to lack structure, since the distances between grouped items appear to be no smaller than the distances between ungrouped items. The distance matrix based on dynamic visemes shows far more structure, with squares of blue forming on the leading diagonal. However, there often appears to be overlap between clusters, which is likely to be because the distances were calculated between the mid-frames of the gestures. These frames represent the transition between two salient poses and are more likely to appear similar across gestures than the entire motion would be.

To further evaluate the efficacy of dynamic visemes for modelling visual speech, a set of objective and subjective tests were performed.

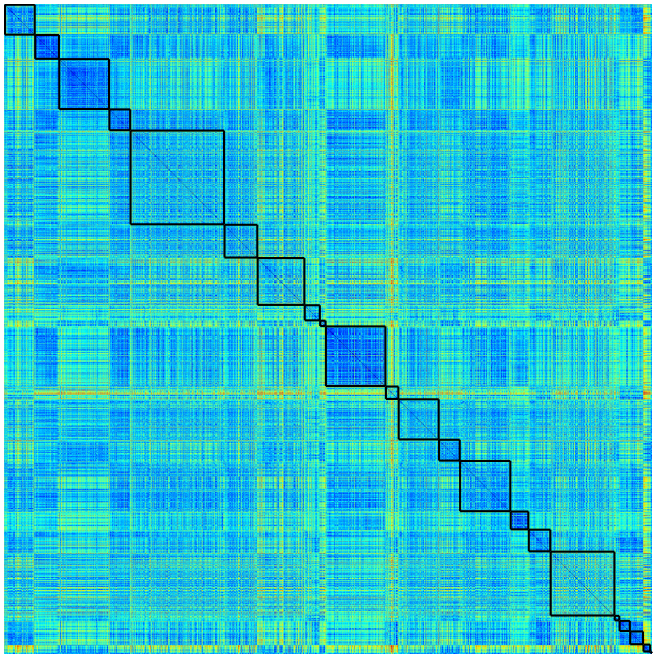


Figure 6.20: Distance matrix showing the Euclidean distance between the combined AAM parameters at the mid-frame of each phone, grouped by the viseme labels as determined by Parke and Waters [121]. The viseme groups are highlighted with the black boxes along the leading diagonal.

6.5.1 Objective Evaluation

For a random set of 500 sentences taken from the clustered data, the AAM trajectories were reanimated using dynamic viseme concatenation and, for comparison, static viseme interpolation. To generate the dynamic viseme trajectories, each of the gestures are simply replaced with the median of the cluster in which it belongs. Where the gestures are of different durations, the median gestures are linearly resampled. The median gesture rather than the mean is chosen as a representative for each dynamic viseme cluster as it does not assume a Gaussian, or symmetric distribution of the gestures within a cluster and is less sensitive to outliers. The median gesture is also desirable as it represents a real trajectory from the training data rather than an approximated one, as is represented by the mean.

The static viseme interpolation method is based on Parke and Waters' eighteen visemes [121] which are presented in Figure 3.1. This set was chosen as it contains a relatively large number of viseme classes and a complete coverage of phonemes,

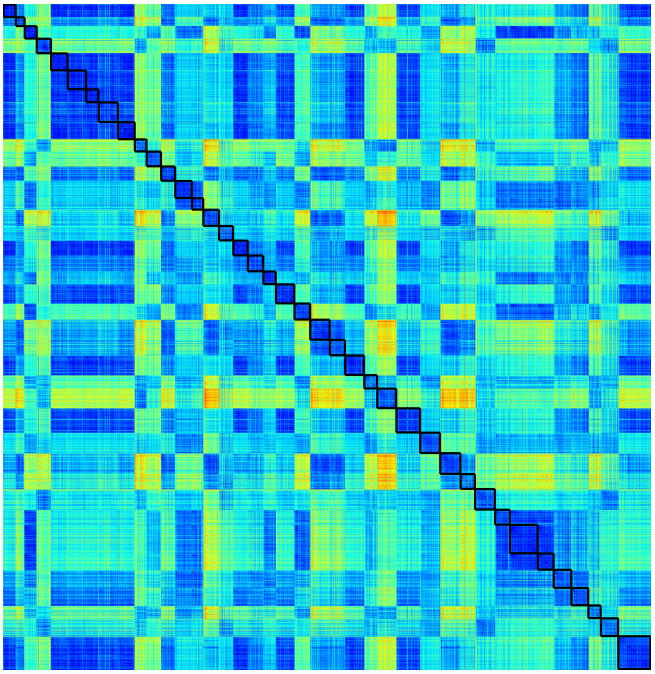


Figure 6.21: Distance matrix showing the Euclidean distance between the combined AAM parameters at the mid-frame of each visual gesture, ordered by the dynamic viseme labels from clustering for the first forty dynamic visemes. The dynamic viseme groups are highlighted with the black boxes along the leading diagonal.

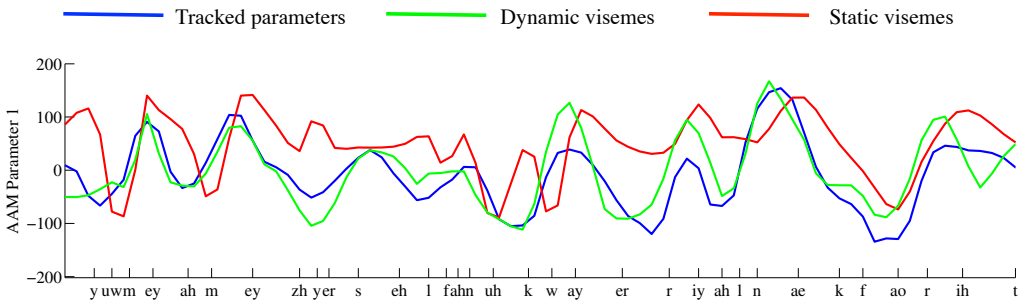
Diphthong	Approximation
/eɪ/	/æy/
/aɪ/	/ʌy/
/ɔɪ/	/ɔy/
/oʊ/	/ʊw/
/aʊ/	/æw/

Table 6.2: Approximating diphthongs with pairs of phonemes.

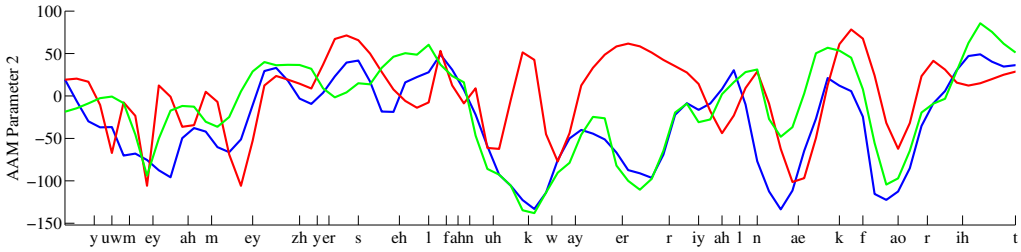
with the exception of diphthongs which are therefore approximated with pairs of phonemes shown in Table 6.2. For each of the static viseme classes, a frame is selected from the KB-2k video containing a pose which closely corresponds to that described and illustrated in [121]. The selected frames are shown in Figure 6.22. The synthesised trajectories are generated by placing a static pose at the mid-frame of each phone segment and interpolating the intermediate frames with a cubic two-dimensional Bezier curve in Autodesk Maya 2011. This simple interpolation method was chosen as it allows direct comparison of static and dynamic viseme *units*, as the synthesised trajectories contain minimal blending in both cases.

Figure 6.22: The mouth region from the selected frames of the KB-2k dataset for each of the 18 visemes as determined by Parke and Waters [121].

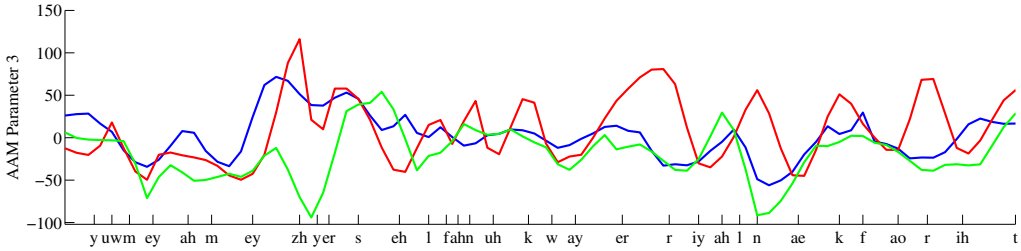
For the sentence "You may amaze yourself and acquire a real knack for it", the synthesised trajectories for the first five components of the AAM generated using dynamic and static visemes are presented in Figure 6.23 with the corresponding trajectories from the tracked video sequences. More examples are shown in Appendix C.1. It is clear that the trajectories generated using the dynamic viseme concatenation method more closely follow the desired trajectories and that the static pose interpolation method appears to be far more erratic as it is constrained to hit each of the targets. To quantify the quality of the generated 20D AAM parameter trajectories, they are compared to the corresponding trajectories measured from the video sequence by calculating the root-mean-square error averaged over the frames from 500 sentences. The mean and standard deviation for each case are shown in Table 6.3. A t-test confirms that the trajectories formed using dynamic visemes are significantly more similar to the tracked parameters than those generated using the static pose interpolation method ($p < 0.001$). However, as the aim of using dynamic visemes for speech animation is to improve the perceptual quality of the speech by making the lip motion appear more natural, a more fitting way to measure the performance of these approaches is by evaluating the visual quality of sequences of speech.



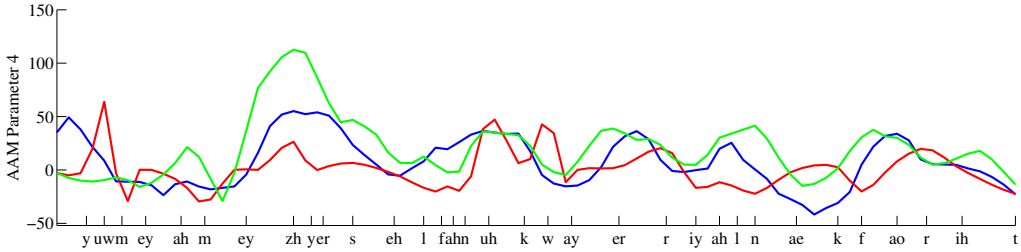
(a) Sentence 156, Parameter 1



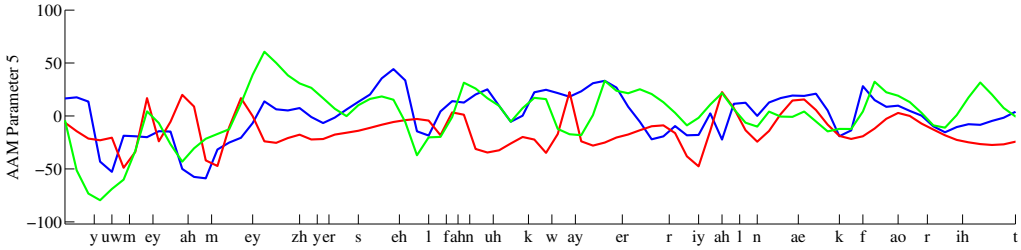
(b) Sentence 156, Parameter 2



(c) Sentence 156, Parameter 3



(d) Sentence 156, Parameter 4



(e) Sentence 156, Parameter 5

Figure 6.23: The temporal trajectories for the first five combined modes of variation of a multi-segment AAM for the sentence "You may amaze yourself and acquire a real knack for it". Shown are the ground-truth parameter values (blue), the concatenated dynamic viseme cluster medians for the known viseme sequences (green) and the interpolated static visemes (red).

	μ	σ
Dynamic visemes:	18.81	8.61
Static visemes:	22.60	8.50

Table 6.3: The mean (μ) and standard deviation (σ) of the RMS error averaged over the frames from 500 sentences for AAM parameters generated both by resynthesising known dynamic viseme sequences and static pose interpolation based on Parke and Waters' eighteen visemes [121].

6.5.2 Subjective Evaluation

To further evaluate the effectiveness of dynamic visemes for modelling visual speech, 50 random sentences from the clustered data are reanimated on a 3D deformer model. Each of the 150 dynamic visemes are animated on the model, and the speech animation is generated by concatenating the known viseme sequences with blending at the join to ensure smooth transitions. For comparison the sentences are also animated based on the phoneme-to-static key pose mapping taken from [121] where keyframes are placed at the mid-point of each phone segment and a cubic two-dimensional Bezier curve in Autodesk Maya 2011 is used to generate the intermediate frames. Again, diphthongs are approximated by concatenating two corresponding vowels as shown in Table 6.2. A description of the deformer model and boundary smoothing, and some examples of both dynamic and static visemes animated on the 3D model can be found in the following chapter. The movies are rendered at a resolution of 1280×720 pixels. The odd animation frames for the sentence "Only rarely is attention given to accurate progress reports and evaluation" generated using dynamic visemes and static pose interpolation are shown in Figures 6.24 and 6.25 respectively. More examples are shown in Appendix C.2.

Thirty two participants took part in a pairwise preference test where, for each sentence, they were shown two movies side-by-side — one for each condition: a) dynamic viseme and b) static pose interpolation. They were played the left movie, followed by the right movie and finally both movies synchronously. After each sentence, viewers selected whether they preferred the left or the right movie. The order of the sequences and the left-right position on screen for each treatment were

Figure 6.24: The odd frames from an animated sequence generated using dynamic visemes for the sentence "Only rarely is attention given to accurate progress reports and evaluation".

Figure 6.25: The odd frames from an animated sequence generated using static pose interpolation for the sentence "Only rarely is attention given to accurate progress reports and evaluation".

randomised for each participant.

The results of this experiment reveal that viewers prefer animation generated using concatenated dynamic visemes to animation using a phoneme-to-static viseme lookup, on average, 80% of the time. A two sided binomial test reveals that this is a significant difference ($p < 0.01$). This result shows that these units are an effective visual analogue of phonemes since a dynamic viseme is always the same example of the unit from the training video, and these are simply concatenated.

6.6 Discussion

A visual speech utterance is described compactly by a trajectory in AAM space. This trajectory is segmented based on the dynamics of the motion of the articulators into a sequence of intuitive, non-overlapping, variable length, dynamic visual speech gestures. All of the gestures from the large KB-2k dataset are then clustered to form a smaller set of groups, which each represent a distinct visual speech action. It is these short actions that should form the building blocks of visual speech, so rather than referring to visemes as the visually contrastive phonemes, instead visemes are redefined as the related **gestures** that are perceived to have the same function visually. In this way visemes serve to represent meaningful contrasts between visual speech utterances.

In this chapter the Euclidean distance, measured in super-feature space was found to most reflect perceptual distances between gestures, and was therefore used as the basis for clustering. A graph-based clustering algorithm grouped the gestures into 150 viseme clusters. Analysis of the clusters shows that there is a complex, many-to-many relationship between phoneme sequences and dynamic visemes, as a phoneme is likely to be distributed over many clusters, and a particular cluster contains many phoneme sequences.

Both objective and subjective evaluation suggest that animation generated using dynamic visemes when the true viseme sequence is known produces significantly

more natural lip motion than using static viseme interpolation. Since dynamic viseme animation was generated by simply concatenating the median gestures from the respective clusters, it suggests that gestures within a viseme cluster portray equivalent visual speech functions, and that the units are effective for modelling visual speech.

Chapter 7

Animating Speech with Dynamic Visemes

In this chapter dynamic visemes are applied to the problem of animating new, unseen visual speech sequences. The task is to transform a series of phoneme labels with corresponding durations to a sequence of dynamic visemes that match the speech sounds. Traditionally animation is done by first substituting the phonemes for their respective *visual phonemes*, then attempting to generate sequences of mouth movements using, say, concatenation [15, 100] or trajectory formation [13, 33, 45]. The difficulty with using a dynamic viseme approach is that phonemes do not map directly to visual labels as they do in a traditional sense, and there are no visemic transcriptions of the spoken words. Consequently a *mapping* is defined from phonemes to dynamic visemes rather than a simple lookup.

This chapter first outlines the various forms of face model that can be animated using dynamic visemes, including an image-based model, a traditional blend shape model and a deformer model. The phoneme-to-dynamic viseme mapping, and the way in which gestures are concatenated for animation is then described. The animated sequences are evaluated using subjective and objective methods, and compared with trajectories formed using traditional phoneme-to-(static) viseme animation. Finally, the speech of a second speaker is segmented and clustered into dy-

namic visemes, which are mapped to the dynamic visemes from the KB-2k dataset. Speaker two is then animated using the viseme sequences from the KB-2k dataset.

7.1 Facial Models for Animation

The AAM allows direct visualisation of animated sequences by reconstructing the image from the parameterisation. Thus, AAM trajectories can be synthesised, and image-based animation can be generated by blending the AAM modelled jaw region onto a background image, for example, using Poisson blending [125] for composition. The addition of non-speech animation, such as expression, is dependent on the implementation of the model.

The AAM parameterisation also allows for animation of any face model that is rigged with blendshapes designed to match the shape eigenvectors, $\mathbf{p}$, of the AAM. The shape parameters can then directly control the lip motion on the model, and although appearance information is lost, this method allows for full 3D rendering and lighting effects. The main disadvantage of this approach is that, as it is a direct mapping of the tracked actor, the visual speech movements and the proportions of the face model are constrained to be somewhat similar to that of the actor's.

The main advantage of dynamic visemes is that they can be used to animate speech on any generic deformer-based model. Since the gestures that form a particular viseme cluster represent the same action on the lips, only one of those gestures must be defined on a character for each cluster. As with traditional, static visemes, dynamic visemes can be artistically modelled, so that different characters can produce the same action, but with a different style, and these only need to be defined once for a given character for any speech utterance to be animated. This means that they are convenient for use in industry, as they can directly replace static visemes which are currently the standard approach to speech animation.

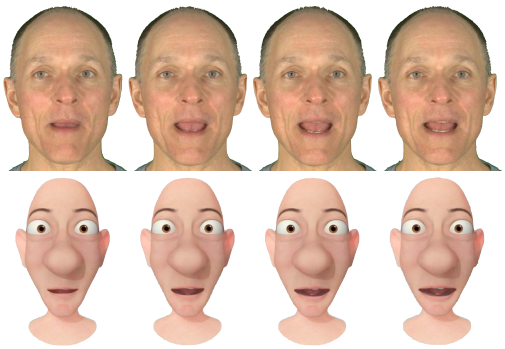
In this chapter speech animation is implemented by animating dynamic visemes on a 3D model that has been artistically rigged using surface deformers in Autodesk

Maya 2011. This represents an industry standard modelling and rigging approach. All gestures belonging to a dynamic viseme serve the same visual function, so each viseme is represented with the median visual gesture of those assigned during clustering, which is then animated on the character. Figure 7.1 shows examples of four dynamic visemes on the speaker, and the corresponding animations on the deformer model. The face model and the dynamic viseme animations were generated by an animator at Disney Research in Pittsburgh, USA.

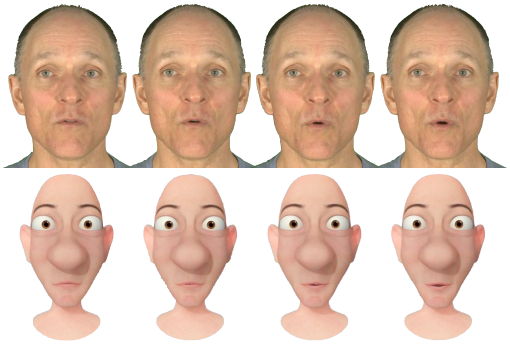
7.2 Mapping Phonemes to Visemes

Given an input sequence of N phoneme labels, $P = p_1, p_2, ..., p_N$, with corresponding durations, an output sequence of M dynamic viseme labels, $V = v_1, v_2, ..., v_M$, that best corresponds to the desired speech movements is required. To find this mapping the phoneme strings that are associated with the viseme clusters during training are exploited. Specifically, each viseme, v_i, has a number of variable length phoneme strings associated with it, corresponding to the constituent gestures assigned during clustering. Using these phoneme strings, an exhaustive search is performed to locate all possible sequences of visemes that could have given rise to the input phoneme sequence P.

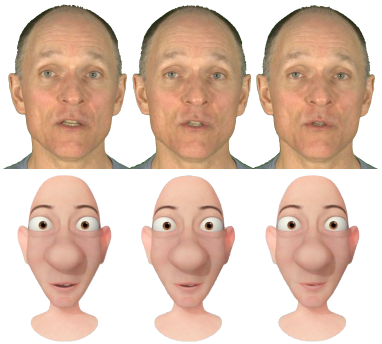
As an example, if the target phrase is "word", the instances of the phoneme string, $P =$ { /w/, /ɜ/, /d/}, are first searched for in the dynamic viseme clusters. Any clusters that contain this sequence are identified as candidate viseme sequences. Next, the phoneme substrings {/w/, /ɜ/} and {/d/} are searched for, and all combinations of dynamic visemes containing these sequences are added to the candidate viseme sequences. Finally, the sequences {/w/} and {/ɜ/, /d/} are searched for. The boundaries between phonemes and dynamic visemes tend not to align, so to account for this, phonemes corresponding to the end of one gesture are also allowed to appear at the beginning of the next gesture during the search. Figure 7.2 illustrates all of the search paths for "word", in which the black nodes represent dynamic viseme



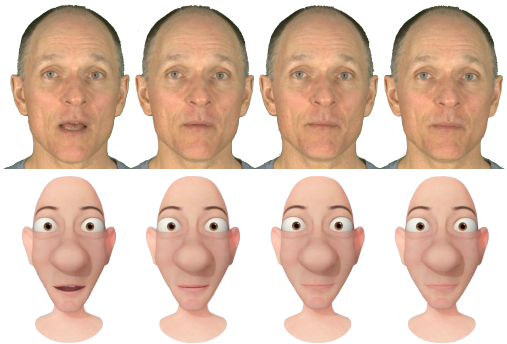
(a) Viseme 5



(b) Viseme 66



(c) Viseme 80



(d) Viseme 90

Figure 7.1: Four example dynamic visemes animated by an artist on a surface-deformer model in Maya.

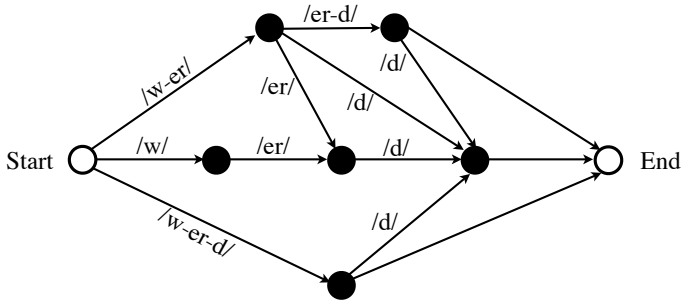


Figure 7.2: Possible paths for mapping the phoneme string /w-ɜ-d/ to visemes (black nodes).

classes, and Figure 7.3 shows a visualisation of a possible path through viseme space for the sentence "Have a listen to this". In this example, a possible dynamic viseme encoding of the phoneme sequence is 33-103-117-98-133-100-40-77-98.

To select the best matching dynamic viseme sequence from a list of candidates, each candidate is assigned a cost as follows:

$$c_i = \alpha(-\Pr(V_i|P)) + \beta(t(V_i, P)) + \gamma(d(V_i)), \tag{7.1}$$

where V_i represents the i^{th} candidate viseme sequence. The first term in Equation 7.1 represents the probability of viseme sequence, V_i, given the phoneme string. This is calculated by summing the bigram log probabilities for the viseme pairs and the log probabilities of the respective phoneme substrings with respect to the viseme cluster:

$$\Pr(V|P) = \sum_{m=2}^{|V|}(\log(\Pr(v_m|v_{m-1}))) + \sum_{m=1}^{|V|}(\log(\Pr(P_m|v_m))). \tag{7.2}$$

The second term in Equation 7.1 represents the cost of temporally aligning the dynamic visemes in V_i to the target sequence P in terms of duration. This term is calculated as the squared difference between the number of frames in the segment of the target sequence and the dynamic viseme, and biases the viseme selection towards those that most closely match the speaking rate of the target sentence.

The final term in Equation 7.1 is a measure of discontinuity at the boundaries of the concatenated dynamic visemes measured in AAM space. This is represented by

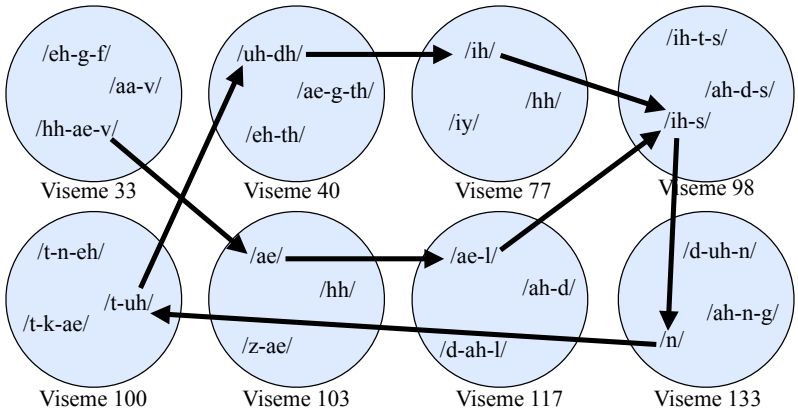


Figure 7.3: Mapping variable length phoneme substrings for the sentence "Have a listen to this" to dynamic visemes.

the Euclidean distance between gesture boundary frames, and biases the selection towards sequences of visemes that generate the smoothest trajectory.

The weights α, β and γ are determined subjectively by visually inspecting the animation generated with different values, and are set to 0.699, 0.3 and 0.001 respectively. The weights vary because the cost terms are measured in different units that have differing ranges. The cost function is far more sensitive to a change in γ than the others as the third cost term is measured in AAM space so typically has larger values. These parameters can be adjusted to vary properties of the output animation, but for all results in this work, these are the values used. On completion of the search algorithm the lowest cost viseme sequence is used to generate the output speech animation by concatenating the corresponding dynamic visemes.

7.2.1 Dynamic Viseme Concatenation

To animate the deformer model, the dynamic visemes in the sequence with the lowest cost are simply concatenated with blending at the boundary frames to create a smooth join. To blend two gestures, the segment start and end frames are replaced with a half-frame, mid-value point. The values for the segment start and end frames are computed using Maya's cubic two-dimensional Bezier curve fitting function to interpolate through the half frame without disrupting other values along

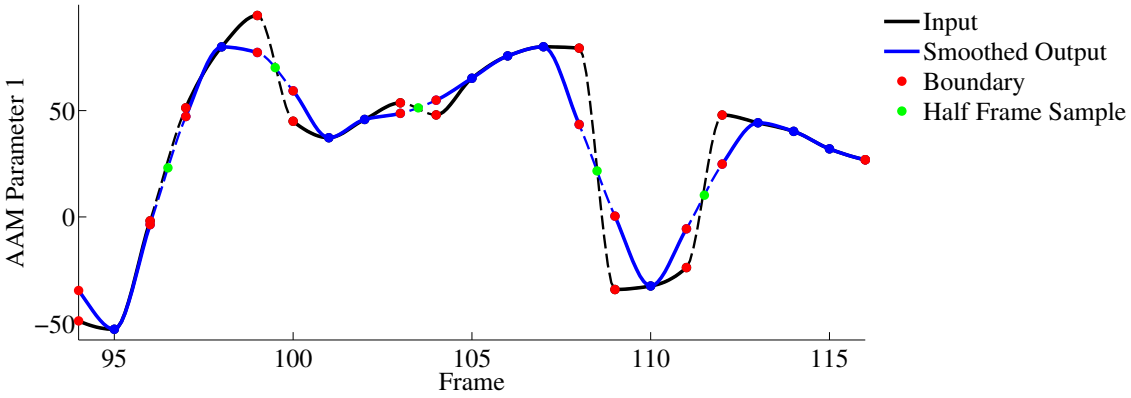


Figure 7.4: To stitch together dynamic visemes, the segment start and end values (black curve, red points) are replaced with a half-frame, mid-value point (green). Default Maya curve interpolation computes new values (blue curve, red points) for the segment start and end values without disrupting other elements along the curve.

the curve. Figure 7.4 shows the effect of the boundary smoothing. Note that only the boundaries are affected by the join, so the viseme dynamics remain the same.

7.2.2 Viseme Alignment

Dynamic visemes are independent of phonemes, so the boundaries tend not to align. However, the viseme boundaries can be approximated from the known phone boundaries using:

$$v_i^e = \begin{cases} p_j^e, & \text{if } v_i^e \text{ does not intersect } p_j \\ \dfrac{p_{j-1}^e + p_j^e}{2}, & \text{otherwise} \end{cases} \tag{7.3}$$

where v_i^e represents the end frame of viseme i, and p_j^e represents the end frame of phoneme p_j. A viseme is assumed to intersect a phoneme if the phoneme label is split over two consecutive visemes, otherwise the boundaries are assumed to align. This exploits the phenomenon that humans do not perceive an offset of 80ms ($\approx$ 3 video frames) when the audio leads and 140 ms ($\approx$ 5 video frames) when the audio lags in speech [140]. As the average gesture length is 110ms, the majority of cases fall within these tolerances.

7.3 Evaluation

In the previous chapter, the efficacy of dynamic visemes for animation was evaluated when the ground-truth viseme sequence was known, as the sentences used were included in the clustering process. In this section, the full animation pipeline is evaluated both objectively and subjectively on a set of 50 sentences that were held out of clustering, so the viseme sequences are unknown and must be generated for the phonemes using the methods described in Section 7.2. This test is designed to measure the quality of the phoneme-to-dynamic viseme lookup and the cost function as well as testing the units.

7.3.1 Objective Evaluation

To quantitively evaluate the quality of the animated sequences, the AAM trajectories are synthesised for the 50 previously held-out test sentences using the phoneme-to-dynamic viseme lookup and viseme alignment described previously. For comparison, the sentences were also generated using the static viseme interpolation approach described in Section 6.5.1. A more advanced coarticulation function was not implemented as, typically, they are defined manually or are data dependent, so it would be unclear how to apply them to a computer-generated model. More importantly, the simple interpolation function allows for a direct comparison of the *units*, as minimal blending is applied in either case.

Figure 7.5 shows the first five AAM parameters of the tracked features, the synthesised trajectories formed by stitching together dynamic viseme sequences and trajectories generated using static pose interpolation. From these graphs it is apparent that the parameters generated using dynamic visemes more closely follow the desired trajectory than the static pose interpolated parameters, as the latter appear amplified and often asynchronous to the desired trajectory. More examples are shown in Appendix D.1.

The quality of the synthesised trajectories is measured by calculating the RMS

	μ	σ
Dynamic visemes:	16.22	$\pm$7.29
Static visemes:	21.87	$\pm$7.11

Table 7.1: The mean (μ) and standard deviation (σ) of the RMS error averaged over the frames from 50 sentences and over all 20 AAM parameters generated both by phoneme-to-dynamic viseme mapping and static pose interpolation based on Parke and Waters' eighteen visemes [121].

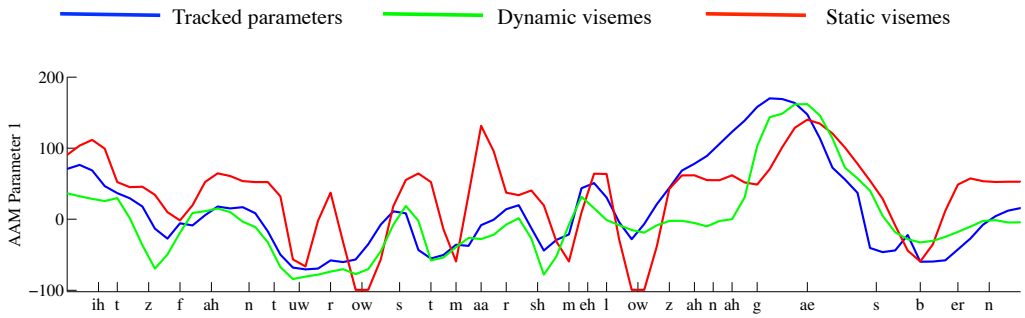
error averaged over all of the frames in the 50 sentences that contain speech. The results are listed in Table 7.1 and show that synthesised trajectories generated using dynamic visemes more closely follow the desired trajectories than those formed using static pose interpolation. A t-test reveals that the benefit is significant ($p < 0.001$).

7.3.2 Subjective Evaluation

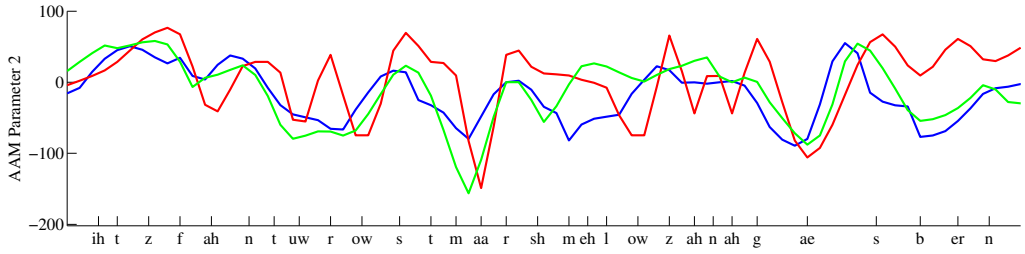
The 50 sentences were animated on the 3D deformer model using the methods described in this chapter. Again, for comparison, Parke and Waters' static visemes [121] were also modelled on the character, and the animated sentences were generated by placing the corresponding visemes at the mid-frame of each phone segment and interpolating between them using a cubic two-dimensional Bezier curve in Autodesk Maya 2011. The eighteen static visemes on the deformer model are shown in Figure 7.6. The animation frames for the sentence "At least the wheels dug in" generated using dynamic visemes and static pose interpolation are shown in Figures 7.7 and 7.8 respectively. More examples are shown in Appendix D.2.

Thirty two participants took part in an experiment which compared animated sequences formed using dynamic and static visemes in the form of a pairwise preference test which followed the same procedure as that described in Section 6.5.2. Viewers again prefer ($p < 0.07$) animation generated using concatenated dynamic visemes to animation using a phoneme-to-static viseme lookup, this time on average, 62% of the time.

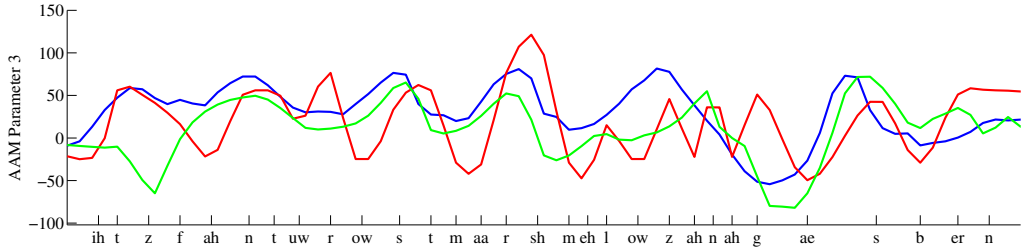
When compared to the viewer preference for the dynamic viseme animations when the viseme sequence is known (80%), it is clear that the dynamic viseme



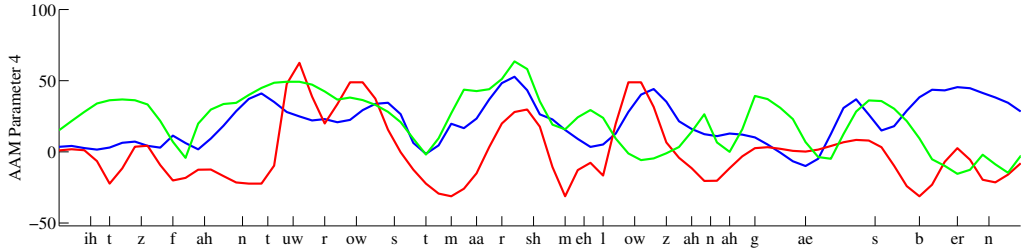
(a) Sentence 1725, Parameter 1



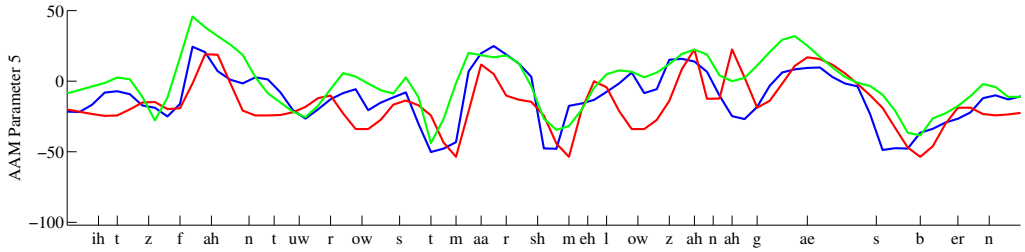
(b) Sentence 1725, Parameter 2



(c) Sentence 1725, Parameter 3



(d) Sentence 1725, Parameter 4



(e) Sentence 1725, Parameter 5

Figure 7.5: The temporal trajectories for the first five combined modes of variation of a multi-segment AAM for the sentence "It's fun to roast marshmallows on a gas burner". Shown are the ground-truth parameter values (blue), the concatenated dynamic viseme cluster medians for the synthesised sequences (green) and the interpolated static visemes (red).

Figure 7.6: The mouth region of the deformer model for each of the 18 visemes as determined by Parke and Waters [121].

lookup process introduces substantial error in the unit selection, and could benefit from further development. Feedback suggests that even a single error in the selection of the animation units can severely impact the perceived quality, so it is important that this is done correctly. However, the results suggest vast potential for use in animation, and further support the suitability of dynamic visemes as the units of visual speech.

7.4 Generalizing Visemes Across Speakers

So far, dynamic visemes have only been considered for a single speaker. However, people speak very differently to one another, so it is possible that different speakers have a different set of dynamic visemes, and that some of the visemes identified using the KB-2k dataset are speaker dependent.

To investigate how well the units generalise across speakers, the speech of a second, female speaker is segmented and clustered into dynamic visemes. This speaker is from the FSpace dataset, recorded at Disney Research, Pittsburgh. The video is recorded at the same frame rate and resolution as the KB-2k dataset, but contains just 200 utterances in contrast to ≈2500. The speech was phonetically annotated, and tracked using AAMs as described in Section 5.2.3. A selection of the

Figure 7.7: Frames from an animated sequence generated using dynamic visemes for the sentence "At least the wheels dug in".

Figure 7.8: Frames from an animated sequence generated using static pose interpolation for the sentence “At least the wheels dug in”.

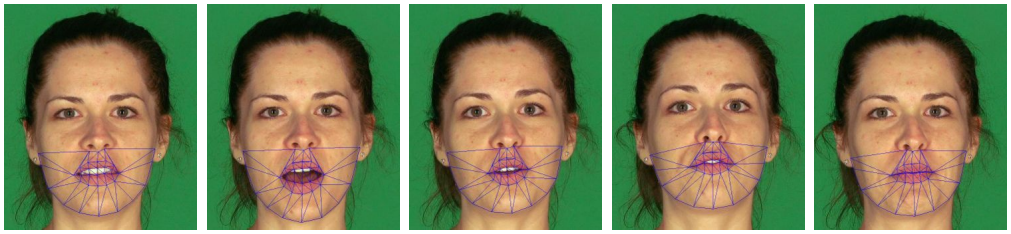


Figure 7.9: A selection of training images used to build the AAM for a second speaker that have been manually annotated with 34 landmarks demarcating the lips and jaw.

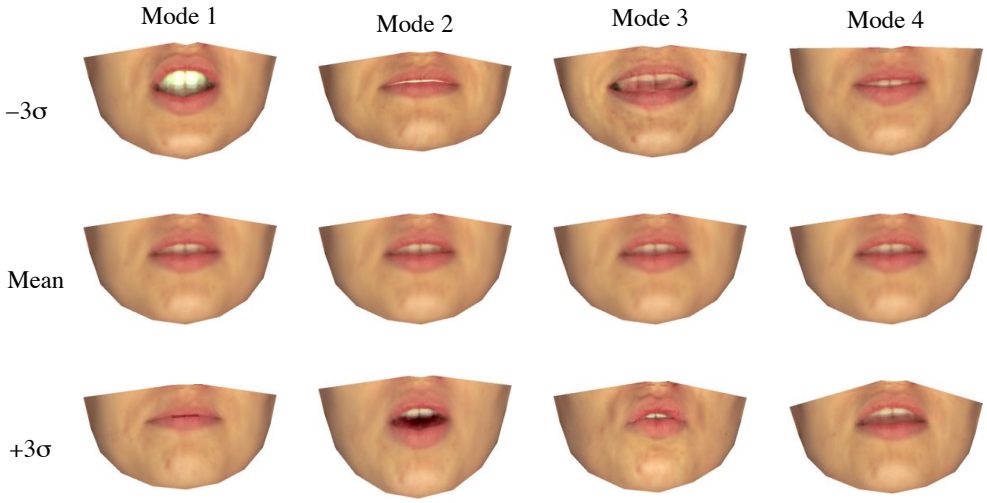
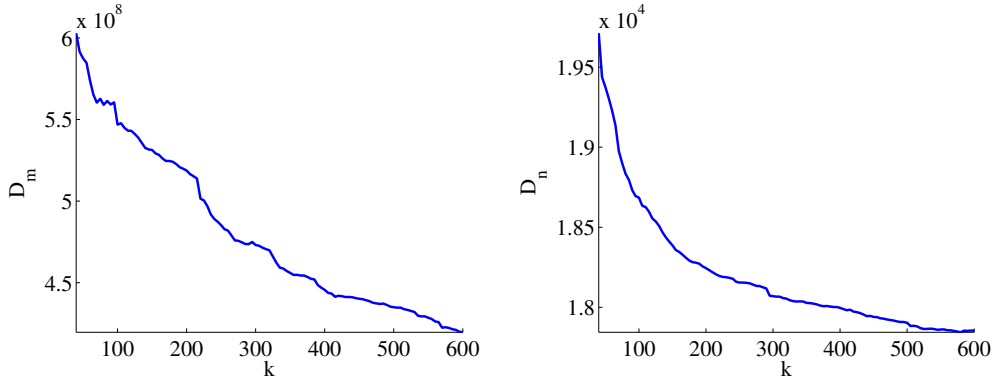


Figure 7.10: Modes of variation for the combined shape and appearance multi-segment model for speaker two at -3 (top) and $+3$ (bottom) standard deviations about the mean (middle).

65 training images that were manually landmarked to train the AAM are presented in Figure 7.9.

The tracked speech is then parameterised using a multi-segment AAM as with KB-2k, where independent appearance models are built for the inner mouth region and the remaining jaw area. For this speaker, 8 shape, 27 jaw appearance and 7 inner mouth appearance parameters describe 95% of the variation. The shape and appearance features for each of the segments are stacked, normalised and PCA is applied to generate a set of 24 features describing the variation in both shape and appearance (see Section 5.3.1). The first four modes of variation of the combined multi-segment model are shown in Figure 7.10.



(a) D_m (b) D_n

Figure 7.11: The mean squared difference between the super-features and the respective cluster median for each gesture (D_m) and the nearest-neighbour from a different cluster (D_n) for speaker 2. The number of clusters is varied over $k = \{40, 45, 50, \ldots, 600\}$. The trade-off value for k is around 150 clusters.

The AAM parameters of the second speaker are automatically segmented into gestures using the methods described in Chapter 6. The gestures are then mapped to speaker dependent super-features, which are clustered. The number of clusters, k, is determined in the same way as speaker one, by calculating the following measures for $k = \{40, 45, 50, \ldots, 600\}$:

D_m the mean distance of the super-features to their respective cluster median.

D_n the mean distance of the super-features to the nearest sample that does not belong to the same cluster.

Figure 7.11 shows the measures as a function of k. Again, to determine the number of clusters required, the knee of the curve is located. There is no clear knee in the curve shown in Figure 7.11(a). However, in Figure 7.11(b) it falls at around 150, which is the same number of dynamic viseme classes determined for Speaker 1.

Comparing the two speaker's viseme clusters is difficult because the AAM parameterisation is speaker specific as the components encode different modes of variation. Therefore the viseme clusters for both speakers are manually corresponded by visual inspection of the cluster median gestures, which defines a mapping between the viseme spaces for the two speakers. A good correspondence is found between the

speech spaces for these talkers. For example, see Figure 7.12 which shows example median gestures from corresponded dynamic visemes for both speakers.

Given the correspondence between talkers, the speech motion from speaker one can be used to drive the speech motion of speaker two. This is done by first using the phoneme to viseme lookup method to estimate the desired viseme sequence from a sequence of input phonemes for speaker one. These visemes are then mapped via the correspondence to speaker two, and the cluster medians are extracted, concatenated and used to drive the face model for the new speaker. Figure 7.13 shows frames taken from the sentence "Draw each graph on a new axis". The real video of speaker one (top) is shown alongside the AAM rendered version of speaker two (bottom). The jaw AAM has been blended onto a neutral, static background using Poisson blending [125].

When viewed with the audio for speaker one, the synthesised sequence appears to be accurately articulating the speech. This means that, given the comprehensive training data for speaker one, a viseme sequence can be generated for an unseen sentence and the speech motions can be transferred directly to a new speaker without the need for recording a full training corpus. All that is required are sufficient examples to estimate the visemes such that the correspondence can be defined.

The results shown here are from a preliminary study whereby the correspondence is defined manually. Further work is necessary to automatically learn the correspondence between viseme clusters for two (or more) speakers.

7.5 Discussion

In this chapter, dynamic visemes were applied to the problem of animating new speech given some phonetically annotated and segmented audio. As the relationship between phone sequences and dynamic visemes is many-to-many and complex, generating viseme sequences for new speech involves a search procedure, as is typical for most other concatenative synthesis approaches. An exhaustive search

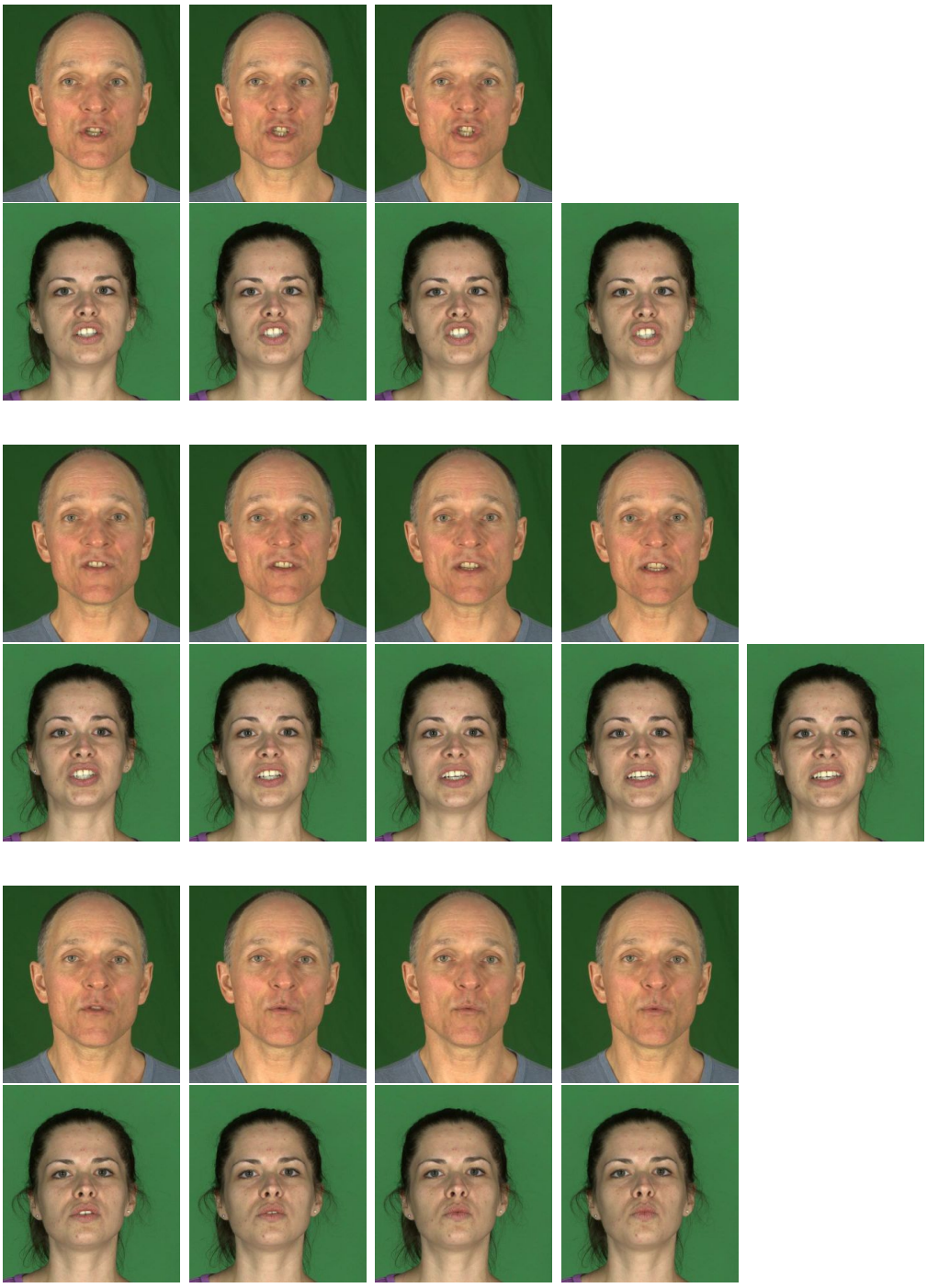


Figure 7.12: Frames from the median gesture of corresponding viseme clusters for two speakers.

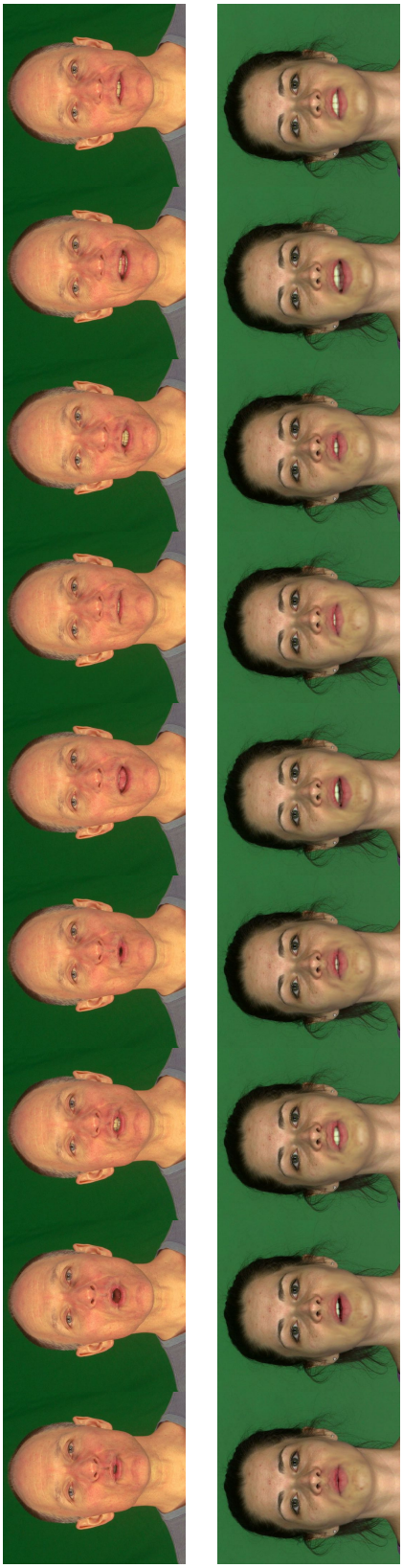


Figure 7.13: Frames taken from the sentence "Draw each graph on a new axis". Here speaker one (top) drives the speech motion of speaker two (bottom) by mapping the viseme spaces.

through viseme space is performed to determine candidate viseme sequences, and the sequence with the lowest cost is selected. The cost is calculated based on the smoothness of the resulting trajectory, the likelihood of the viseme sequences, and the difference between the speaking rate of the viseme sequence and target sentence. To generate animation, the median gestures of the dynamic viseme clusters are modelled on a 3D character, and are simply stitched together in the order specified by the best viseme path with simple blending at the boundary frames.

A subjective and objective evaluation shows that speech animation using dynamic visemes appears significantly more natural and plausible than using static viseme interpolation, although some error in the unit selection is introduced during the viseme lookup process. These results suggest that speech animation generated using dynamic visemes would require less artistic adaptation from the animator. However, the viseme lookup process could benefit from further development.

Finally, visemes from one speaker were manually corresponded with those from a different speaker with an order of magnitude less training data. A good correspondence was determined, and preliminary results suggest that it is possible to animate a new speaker by calculating the viseme sequence using the original speaker's clusters and concatenating the mapped sequence on the new speaker. However, further research is necessary to automate the viseme correspondence, and formally evaluate the viability of the approach.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

In this thesis a new, dynamic unit for visual speech has been presented. Traditionally, static visemes are defined as the groups of visually confusable phonemes, whereas dynamic visemes describe a set of visually contrastive speech movements. Dynamic visemes are derived from an analysis of visual speech rather than by clustering based on phoneme labels, so they represent a true visual analogue to the phonemes of acoustic speech.

Dynamic visemes are identified by segmenting a large corpus of video speech data into sequences of short, non-overlapping movements, which are referred to as visual speech gestures. An audio-visual speech dataset is compactly modelled using an active appearance model and the gesture segment boundaries are defined as the points in parameter space where the acceleration changes from negative to positive. These gestures are clustered to generate a set of 150 reliable and visually intuitive dynamic visemes, where the speech gestures that appear within a viseme cluster all have the same visual function. These units explicitly model coarticulation, maintain the dynamics of the training data, and can be concatenated with simple blending at the joins to animate speech.

Dynamic visemes can and often do span more than one phoneme, and they have a complex many-to-many relationship with phoneme sequences. To generate animation for new speech, any phoneme sequence can be mapped to a sequence of dynamic visemes by exhaustively searching the graph of viseme transitions to find candidate sequences that match the phoneme string. The best viseme sequence is then selected by minimising a cost which is calculated based on the smoothness of the resulting trajectory, the likelihood of the viseme sequences, and the difference between the speaking rate of the viseme sequence and target sentence.

An advantage of using dynamic visemes for speech animation is that they are applicable to any form of facial model or rigging, unlike other concatenative methods which are sample-based and specific to the model on which they are trained. For example, the parameterisation allows direct viewing of the model used for analysis using an image-based renderer or a 3D blendshape model which is rigged with equivalent parameterisation. Alternatively, dynamic visemes are able to drive a completely artistically defined 3D surface deformer model. To do this, an artist must define short animation clips that represent an example gesture of each of the dynamic visemes. These are short clips of mouth movements, typically four or five frames, that only need be defined once for each character.

Using a deformer model, a synthesiser that stitches together dynamic visemes using simple spline curve interpolation at the unit boundaries was demonstrated. This method was compared to a traditional static pose interpolation approach by calculating the RMS error between the synthesised and tracked trajectories in AAM space. The error for dynamic visemes was lower than the error for the more traditional approach. Additionally, in a subjective experiment, participants preferred animation generated using dynamic visemes over static pose interpolation. This indicates that dynamic visemes create more natural and plausible animation, and therefore function as a better foundation for animators than more traditional methods.

Preliminary work on corresponding one speaker's dynamic visemes to another

speaker was presented. Based on a mapping defined by manually corresponding dynamic viseme classes across two speakers, speech animation was successfully transferred from one speaker to another simply by replaying the mapped visemes. A good correspondence was found between the two speakers' dynamic visemes, indicating that the units are at least partially speaker independent. However, further investigation is necessary to automate the correspondence so that the viseme classes across a larger number of speakers can be compared.

8.2 Future Work

This is the first time that dynamic visemes have been presented, and there are a number of further developments that could potentially improve and augment their use for speech animation. For example, although preliminary work on speaker independent dynamic visemes was presented in Section 7.4, it remains unclear which of the dynamic visemes are generic across speakers and whether there are any that are specific to the speaker in the KB-2k dataset. It would be necessary to capture speech from a large number of speakers to determine either a speaker independent parameterisation, or a function to automatically map from one speaker's AAM space to another's so that the similarity between different speakers' gestures can be calculated numerically. A speaker independent clustering can then be performed, and dynamic viseme classes that are speaker specific can be ignored, or used sparingly for animation.

This section outlines further considerations that could refine dynamic visemes for speech animation.

8.2.1 3D Dynamic Visemes

Ijsseldijk [70] measured the speech-reading accuracy of a person speaking when the face was presented at different angles. He discovered that speech-reading accuracy was higher when participants were presented with speech at a frontal view followed

by a repetition at a 60 degree angle, rather than two repetitions at a frontal view. This suggests that the profile view of a person speaking contains complimentary information to the frontal view. Therefore, dynamic visemes might be enhanced by learning from 3D speech data. Additionally, 3D visemes could provide a better reference for modelling dynamic visemes on a 3D computer generated character, as depth information would be available to the animator.

8.2.2 Prosody

Intonation and lexical stress are overlooked in this work. However, stressed phones are typically louder, have a higher pitch and last longer than unstressed phones and are produced with larger facial deformations [149]. Additionally, it has been shown that stressed phones account for more of the coarticulation effects in speech than unstressed phones [4], so it is important to consider the effect of stress to generate realistic speech animation. Dynamic visemes might benefit from training on a dataset containing labels specifying the stressed syllables, as the relationship between stress and the visual gestures could be modelled. To animate new speech, the phoneme sequence must also be annotated with stress labels, but the animation pipeline would remain the same.

8.2.3 Expression

Expression is an important component of visual speech as it helps to convey the emotion of a speaker. Currently, when using dynamic visemes for speech animation, expression must be added as a post process by an animator. It is unlikely that emotional speech is simply a linear combination of neutral speech and facial expression. Rather, the expression changes the dynamics of the visible articulators in a complex way. It is therefore desirable to model expression within the viseme units. One way of accomplishing this is by re-learning dynamic visemes on a dataset of emotional speech. A potential problem is that if many more dynamic visemes

classes are determined this way, dynamic visemes may be less attractive for producing short animations, as the time necessary for modelling the visemes on each character would have increased. However, for feature length animated movies, or video games where a large amount of speech is required, emotional dynamic visemes would remain a viable solution for realistic facial animation.

An alternative approach is to learn whether the underlying sequence of existing dynamic visemes can be altered to generate emotional speech. This way, the number of dynamic viseme classes remains the same, as only the order in which they are concatenated for animation differs as a function of expression.

8.2.4 Speaking Rate

In a preliminary study, the effect of speaking rate on speech production was measured on a dataset containing an actor speaking 10 sentences from the TIMIT sentence list [117] at 3 speeds (slow, normal and fast), each repeated 10 times ($10 \times 3 \times 10 = 300$ utterances). The prompts were presented in a randomised order in which the sentences and speaking rates were varied. The speaking rates of the uttered sentences are generally in accordance with the speeds that the actor was asked to speak (see Figure 8.1).

The speech was phonetically labelled, and the effect of speaking rate on acoustic speech is measured by calculating the Levenshtein distance [93] between the phone sequences that were uttered during repetitions of the same sentences, spoken at different speeds. The phonetic transcription for each sentence was aligned with all other repetitions of that sentence using forced alignment in HTK [159]. The similarity between the aligned sentences is calculated using the accuracy measure used widely in speech recognition:

$$similarity = \frac{N - D - S - I}{N} \times 100, \tag{8.1}$$

where N is the total number of labels in the reference sentences, D is the number

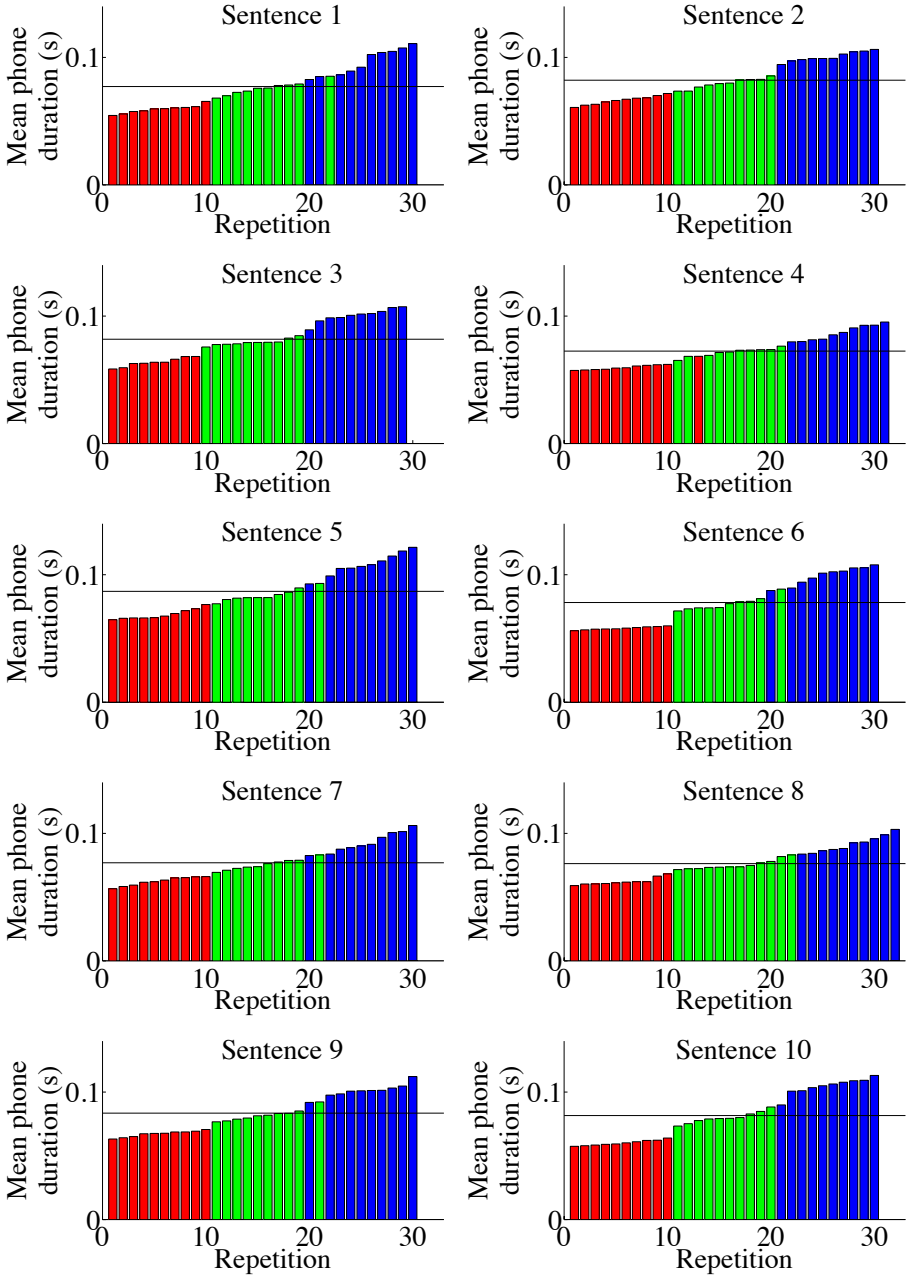


Figure 8.1: The mean phone duration in seconds for each repetition of the 10 sentences in the KB-extra dataset. The colours represent the speed that the actor was asked to speak, where red represents fast, green represents normal and blue represents fast speech.

Spoken	Slow	Normal	Fast
Slow	89.2	87.5	84.8
Normal	86.8	90.7	88.9
Fast	83.3	88.5	88.6

Table 8.1: The phonemic similarity of speech sequences spoken at different rates (slow, medium and fast).

of deleted phones, S is the number of substituted phones, and I is the number of inserted phones. The similarity is averaged over the 10 repetitions for each of the 3 speaking rates, and the results are shown in Table 8.1. The results indicate that even when speaking the same sentence at the same speaking rate, the phones uttered are not entirely identical. The lowest similarity is measured between sentences that are spoken at a fast rate and those that are spoken slowly. A selection of aligned phoneme sequences for the sentence "Almonds and pistachio nuts are not so high in oil, but are rich in protein" is shown in Table 8.2.

To measure the effect of speaking rate on *visual* speech, the movement of the jaw and lips was parameterised using an AAM, as described in Sections 5.2.3. The speech was automatically segmented into gestures based on the zero crossings in acceleration (see Section 6.2), and the gestures were clustered into 150 dynamic viseme classes, providing a dynamic viseme transcription of the sentences. As with the phoneme labels, for each sentence the viseme labels were aligned with all other repetitions of the sentence using forced alignment, and the similarity was measured using Equation 8.1. The results are shown in Table 8.3. Note the negative value comparing slow and fast speech, which suggests that a large number of speech units present in slow speech are missing from fast speech, and that speech that is spoken at a faster rate is not simply equivalent to slow speech that has been sped up. This is confirmed in Table 8.2 (bottom), which shows a selection of viseme sequences for a sentence, which have been aligned for visualisation. In all cases, the viseme sequences for sentences uttered at a particular speaking rate are more similar to others spoken at the same speed than those that are spoken at different speeds, and the faster the sentence is produced, the fewer visemes are used. Furthermore, the

Slow	
ɑ l m ʌ n	z - æ n p ɪ s t æ ʃ i oʊ n ɑ t s - ɑ r n ɑ t s oʊ h aɪ ɪ n ɔɪ l - b ʌ t ɑ r ɪ tʃ - ɪ n p r oʊ t i n
ɑ l m ʌ n	z - æ n d p ɪ s t æ ʃ i oʊ n ʌ t s - ɑ n ɑ t s oʊ h aɪ ɪ n ɔɪ l - b ʌ t ɑ r ɪ tʃ - ɪ n p r oʊ t i n
Normal	
ɑ l m ʌ n d z	ʌ n p ɪ s t æ ʃ y uh n ɑ t s ɑ r n ɑ t s oʊ h aɪ ɪ n ɔɪ l b ʌ d ɑ r ɪ tʃ ɪ n p r oʊ t i n
ɑ l m ʌ n	z ʌ n p ɪ s t æ ʃ y ʌ n ɑ t s ɑ r n ɑ t s oʊ h aɪ ɪ n ɔɪ l b ʌ d ɑ r ɪ tʃ ɪ n p r oʊ t i n
Fast	
ɑ l m ʌ n	z ʌ n p ɪ s t æ ʃ oʊ n ʌ t s ɑ n ɑ t s oʊ h aɪ ɪ n ɔɪ l b ʌ t ɑ r ɪ tʃ ɪ n p r oʊ t i n
ɑ l m ʌ n	z n p ʌ s t æ ʃ oʊ n ʌ t s ɑ r n ɑ t s oʊ h aɪ ɪ n ɔɪ l b ʌ d ɑ r ɪ tʃ ɪ n p r oʊ t i n

Slow	
5 56 67 11 83 56 62 54 93 97 92 43 99 27	64 50 55 13 49 58 98 53 88 19 99 74 6 89 93 35 45 86 79 41
5 56 67 11 83 56 62 54 93 97 66 37 84 29 74	64 72 101 55 13 49 63 58 98 44 19 27 6 80 22 84 18 35 45 86 71 73
Normal	
5 56 61 56 62 54 100 80 66 67 84 87 91	55 13 49 58 98 84 88 19 38 80 93 35 45 86 71 52
5 56 61 56 62 54 100 80 87 67 84 87 0	13 49 58 98 44 88 38 89 93 35 101 86 71 52
Fast	
5 75 8 62 83 80 87 84 87 0 83 80 70	88 6 89 93 35 81 71 52
5 96 61 8 62 83 80 87 84 87 0 83 58 70	88 100 89 61 35 81 79 41

Table 8.2: A selection of aligned phoneme sequences (top) and gesture sequences (bottom) for the sentence "Almonds and pistachio nuts are not so high in oil, but are rich in protein" spoken at different rates. A dash (-) denotes a short pause.

Spoken		Slow	Normal	Fast
	Slow	53.9	45.5	27.4
	Normal	32.3	57.5	39.8
	Fast	-11.9	23.7	57.6

Table 8.3: The visemic similarity of speech sequences spoken at different rates (slow, medium and fast). Note the negative value comparing slow and fast speech, which suggests that a large number of speech units present in slow speech are missing from fast speech.

results indicate that visual speech is far more influenced by the effect of variable speaking rate than acoustic speech as the difference in similarity is larger and the number of units in a sequence is more variable.

Although speaking rate is accounted for in the cost function (Equation 7.1), the KB-2k dataset only contains speech uttered at a normal speaking rate, so animation generated for fast or slow speech may be suboptimal. It is therefore likely that re-learning the dynamic visemes on a dataset that includes variable speed speech will enhance the phoneme-to-dynamic viseme lookup and produce more natural speech animation.

8.2.5 Model Independence

In this work, dynamic visemes are shown to be effective for animating speech on a 3D surface deformer model. However, by their nature, dynamic visemes are applicable to a variety of facial models. For example, a second 3D model is shown in the middle row of Figure 8.2 which is also implemented in Autodesk Maya 2011 but uses linear blendshape rigging where the blendshapes were designed to match the shape eigenvectors, $\mathbf{p}$, of the AAM. However, since $\mathbf{p}$ were calculated from 2D images, difficulties arose when mapping to 3D blendshapes, and artefacts are apparent during lip rounding. Furthermore, since only the shape information is used to drive the articulators, information regarding the visibility of the teeth and tongue is lost.

The bottom row of Figure 8.2 shows an image-based renderer where the jaw image is reconstructed from the AAM parameterisation and blended onto a static back-

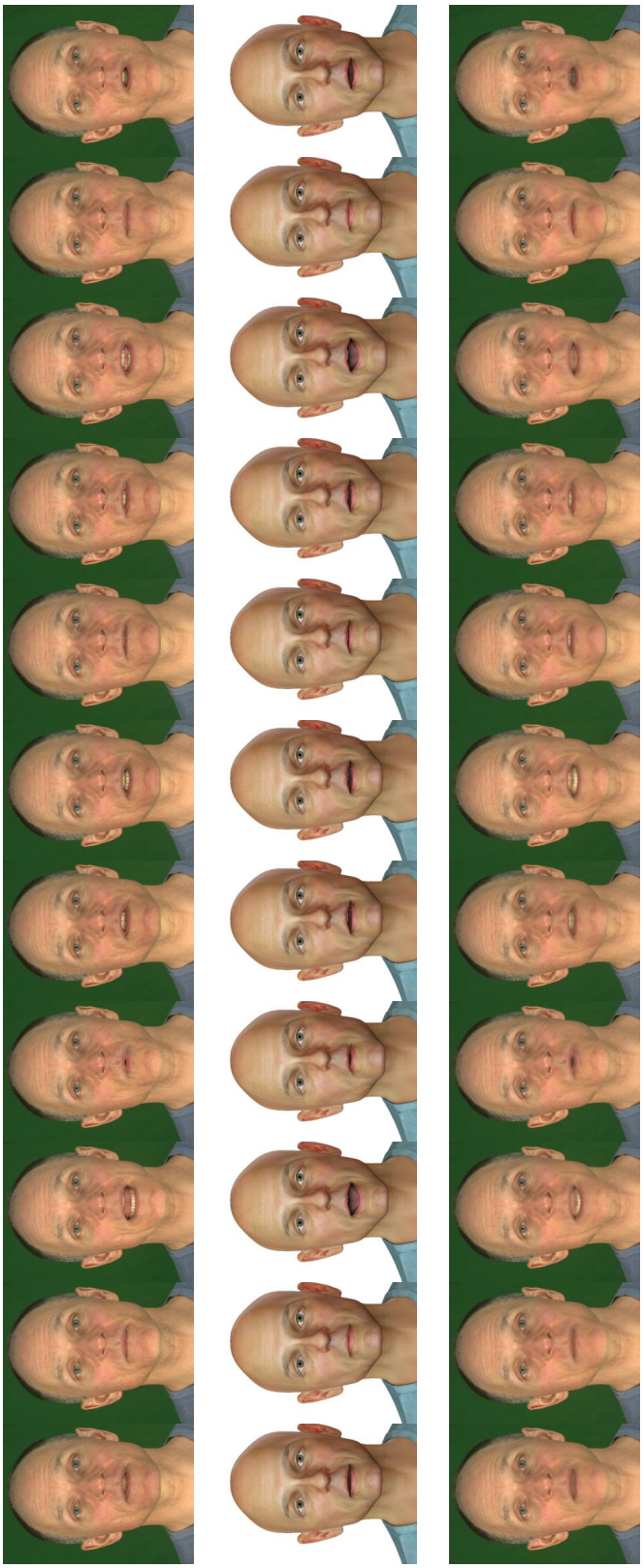


Figure 8.2: Example frames from an animation sequence for a blendshape model designed to match the eigenvectors of the AAM analysis model for direct parameter mapping (middle row), and an image based render that places a Poisson blended AAM synthesis into a static image (bottom row). The original movie frames are shown on the top row.

ground image using Poisson blending [125] for composition. Artefacts are common using this renderer as the boundary between the neck and lower jaw often blurs. The static background is also distracting since it appears unnatural.

In both cases, the dynamics of the articulators appears natural and realistic. However, further work is necessary to improve the quality of the renderers, and to formally evaluate their performance.

8.2.6 Implementation Issues

Currently, no effort has been made to optimise the efficiency of the phoneme-to-dynamic viseme lookup, and the search has an exponential time complexity, taking approximately 4 minutes per sentence. This is still significantly less time than it would take to manually animate the equivalent sequences. However, this duration could be significantly reduced by organising the phonemes in such a way to generate a tree structure, which is far quicker to search. Alternatively, an algorithm such as Viterbi [50] could be used, rather than performing an exhaustive search.

The quality of animation is highly dependent on the phonetic segmentation of both the training data and the sentence to be animated, as the phoneme to viseme lookup exploits the phoneme labels and durations in the cost function. Currently, phonetic annotation is performed manually, so is prone to misalignment and natural variation. The system may be improved by automating this process using an acoustic speech recogniser.

Appendix A

Principal Components Analysis

Principal components analysis (PCA) is a mathematical procedure which can be used for dimensionality reduction, and for learning the structure of data in terms of variation. It transforms a set of n possibly correlated variables to a typically smaller set of orthogonal, un-correlated variables, such that the largest variation of the data is captured in the first axis, the second highest is captured in the second axis, and so on. These new axes are known as the *principal components* of the data. Data reduction is performed by projecting the original coordinates onto the basis formed from PCA, and ignoring higher modes that account for only a small amount variability in the data. The principal components for a set of trivariate data is illustrated is Figure A.1.

There are various methods for calculating the principal components of a set of data. In the work described in this thesis, the eigenvalue decomposition approach is used [157]. The data is represented as a matrix, $\mathbf{X}$, of dimension $m \times n$, where m is the number of observations and n is the number of variables. The mean of each column is subtracted such that the observations are zero centred, and the covariance matrix is calculated:

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X} \tag{A.1}$$

An eigenvector, $\mathbf{u_i}$ and eigenvalue, λ_i, of the matrix, $\mathbf{C}$ satisfies the following

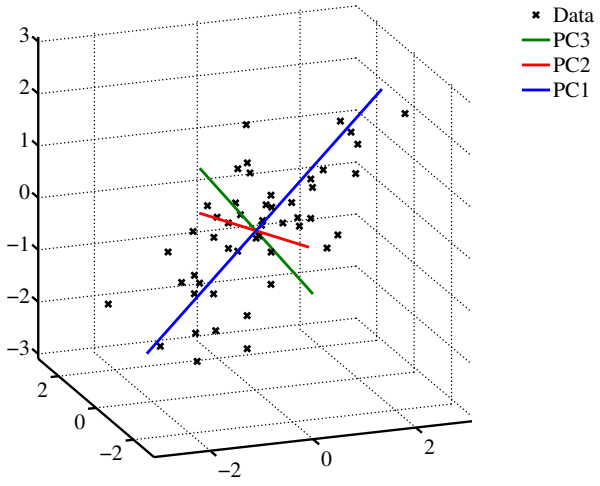


Figure A.1: Principal components analysis of trivariate data with a normal distribution, centred at zero. The eigenvector that explains the most variation, the principal component, is shown in blue, the second principal component is shown in red and the third in green. For data compression, the original coordinates are projected onto the orthogonal basis defined by the eigenvectors, and higher modes are ignored.

linear equation:

$$\mathbf{C}\mathbf{u}_i = \lambda_i \mathbf{u}_i \tag{A.2}$$

An eigenvector of the matrix $\mathbf{C}$ has the quality that only the length, and not the direction changes when multiplying by $\mathbf{C}$. The extent of the change in length is represented by the eigenvalue. The eigenvectors can be represented as a matrix $\mathbf{U}$, such that each column of $\mathbf{U}$ represents an eigenvector of $\mathbf{C}$:

$$\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n), \tag{A.3}$$

and the eigenvectors are stored in the diagonal elements of the matrix, $\mathbf{\Lambda}$:

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & 0 & \ldots \\ 0 & \lambda_2 & 0 & \ldots \\ & & \ddots & \\ 0 & 0 & 0 & \lambda_n \end{pmatrix} \tag{A.4}$$

Equation A.2 can be now rewritten in matrix form:

$$\mathbf{CU} = \mathbf{\Lambda U} \tag{A.5}$$

which, when rearranged, becomes[1]:

$$\mathbf{C} = \mathbf{U\Lambda U}^{-1} = \mathbf{U\Lambda U}^{T} \tag{A.6}$$

The eigenvectors and eigenvalues are then rearranged in order of decreasing eigenvalue. Since the eigenvalues explain the variance in each of the eigenvectors, this step orders $\mathbf{U}$ such that the first column represents the principal mode of variation of the data, the second column represents the second principal mode of variation, and so on. Dimensionality reduction is performed by projecting $\mathbf{X}$ on to the reduced set of basis vectors, $\mathbf{P}$, formed from the first k eigenvectors, where $1 < k < n$.

$$\mathbf{b}_i = (\mathbf{x}_i - \bar{\mathbf{x}})\mathbf{P} \tag{A.7}$$

The number of modes in the projection matrix, k, is chosen so that the required percentage of variance in the original data is preserved. The higher principal components typically model small variations in the data that can be interpreted as noise, and are not important for analysis.

[1]In this case $\mathbf{U\Lambda U}^{-1}$ and $\mathbf{U\Lambda U}^{T}$ are equivalent as covariance matrices are always symmetric and positive definite.

Appendix B

Distance Functions

There is a large number of ways to measure the distance between two vectors, $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ and $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$. The following were implemented for comparison:

1. Euclidean distance

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{B.1}$$

2. Sum of squared error (SSE)

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} (x_i - y_i)^2 \tag{B.2}$$

3. Minkowski distance

$$d(\mathbf{x}, \mathbf{y}) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}} \tag{B.3}$$

   where $p > 0$

4. Manhattan distance

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i| \tag{B.4}$$

5. Cosine distance

$$d(\mathbf{x}, \mathbf{y}) = -\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \tag{B.5}$$

6. Chi squared distance

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{x_i + y_i} \tag{B.6}$$

7. Canberra distance

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|} \tag{B.7}$$

8. Mahalanobis distance

$$d(\mathbf{x}, \mathbf{y}) = -\sum_{i=1}^{n} z_i x_i y_i \tag{B.8}$$

where $z_i = \sqrt{1/\lambda_i}$ and λ_i is the eigenvalue for the i^{th} principle component. This is also true for equations (B.9), (B.10) and (B.11)

9. Normalised mahalanobis distance

$$d(\mathbf{x}, \mathbf{y}) = -\frac{1}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \sum_{i=1}^{n} z_i x_i y_i \tag{B.9}$$

10. Weighted manhattan distance

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} z_i \, |x_i - y_i| \tag{B.10}$$

11. Weighted SSE

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} z_i (x_i - y_i)^2 \tag{B.11}$$

12. Modified SSE

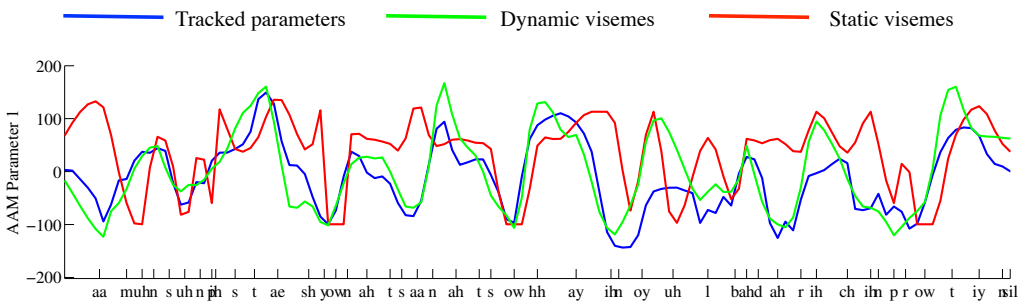
$$d(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n} (x_i - y_i)^2}{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2} \tag{B.12}$$

Appendix C

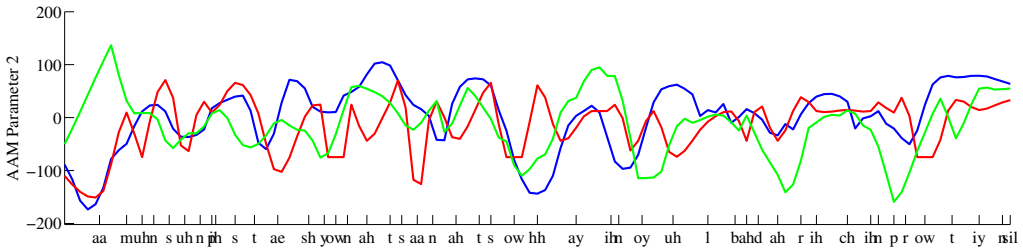
Animation Output for Training Sentences

C.1 Trajectories

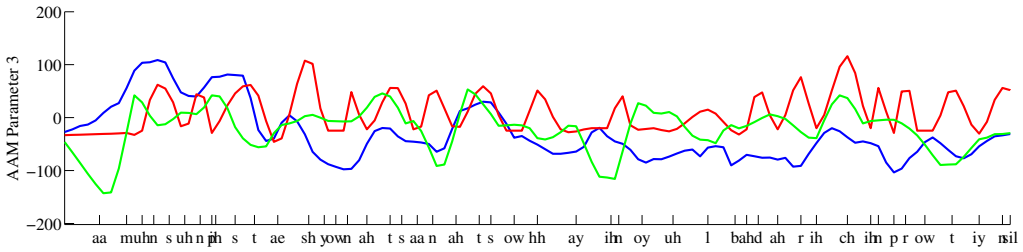
This section contains a selection of synthesised AAM trajectories generated by concatenating known dynamic viseme sequences with blending at the boundary frames as described in Section 6.5.1. For comparison, the true AAM parameters and the trajectories formed from a traditional phoneme-to-static viseme mapping are presented.



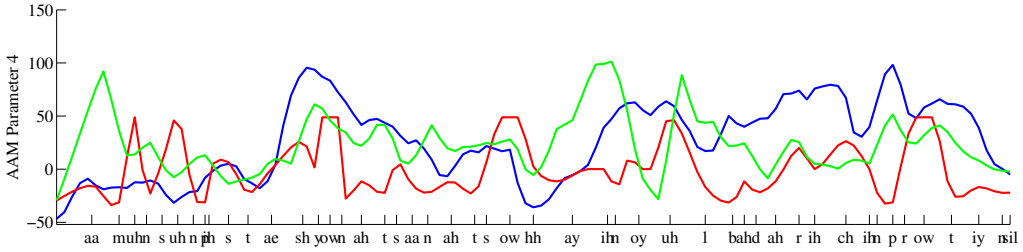
(a) Sentence 5, Parameter 1



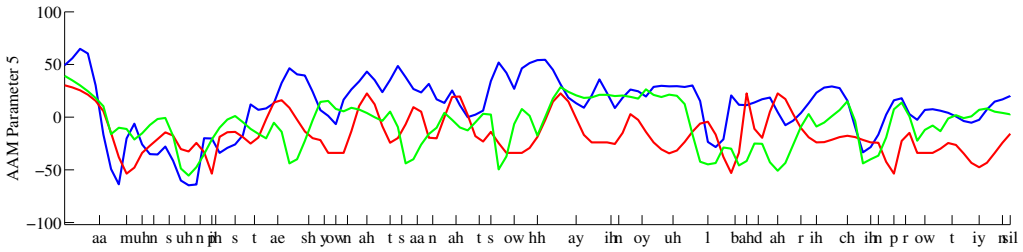
(b) Sentence 5, Parameter 2



(c) Sentence 5, Parameter 3

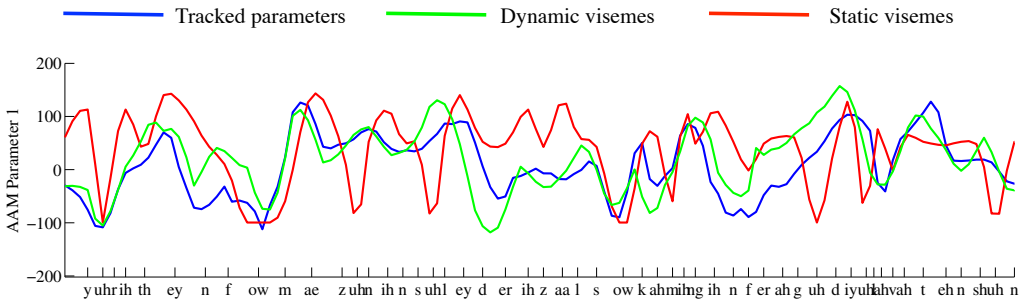


(d) Sentence 5, Parameter 4

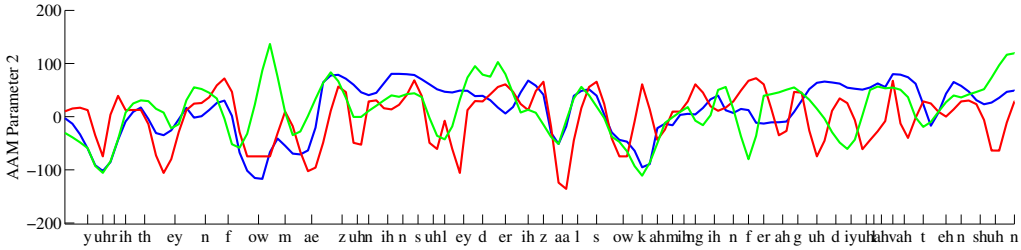


(e) Sentence 5, Parameter 5

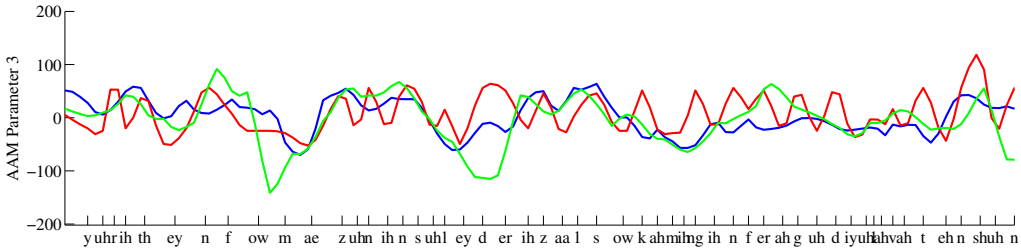
Figure C.1: The temporal trajectories for the first five combined modes of variation of a multi-segment AAM for the sentence "Almonds and pistachio nuts are not so high in oil, but are rich in protein". Shown are the ground-truth parameter values (blue), the concatenated dynamic viseme cluster medians for the known viseme sequences (green) and the interpolated static visemes (red).



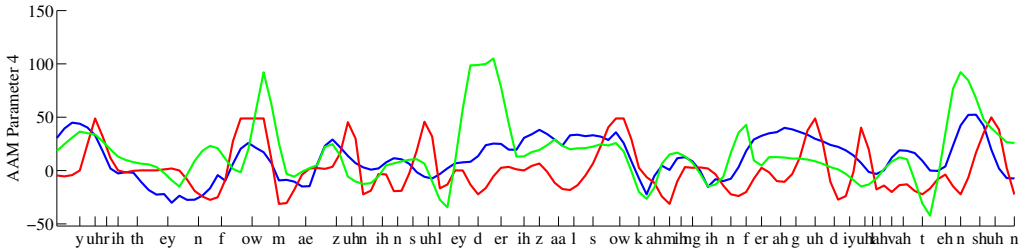
(a) Sentence 59, Parameter 1



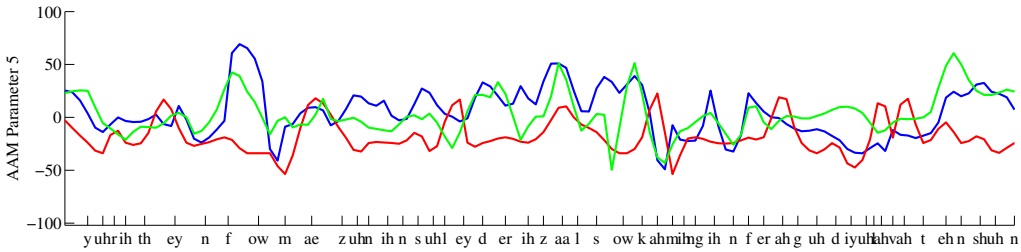
(b) Sentence 59, Parameter 2



(c) Sentence 59, Parameter 3

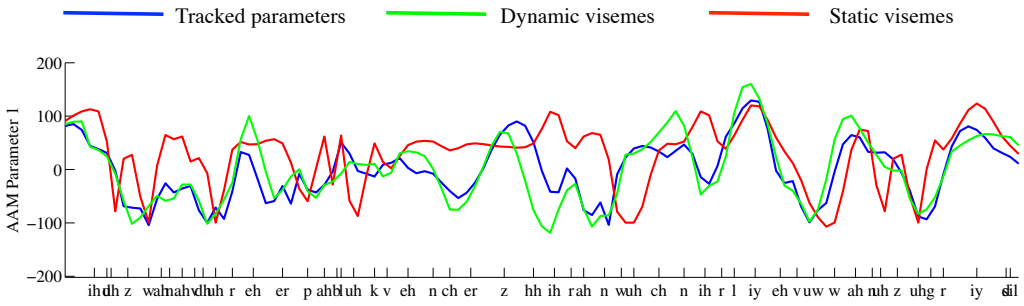


(d) Sentence 59, Parameter 4

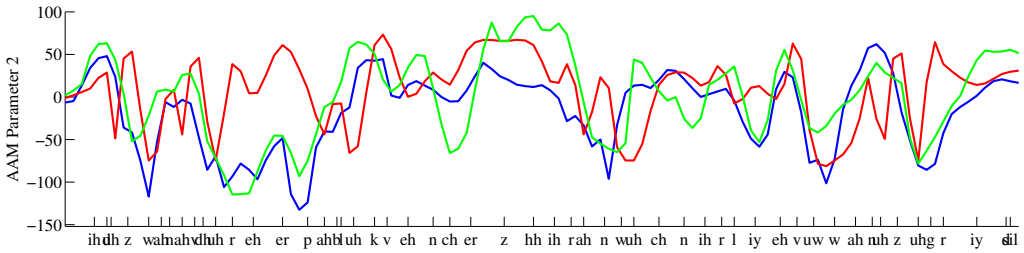


(e) Sentence 59, Parameter 5

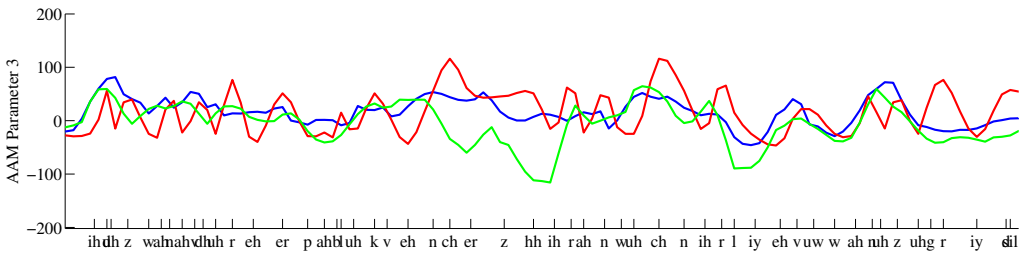
Figure C.2: The temporal trajectories for the first five combined modes of variation of a multi-segment AAM for the sentence "Urethane foam as an insulator is also coming in for a good deal of attention".



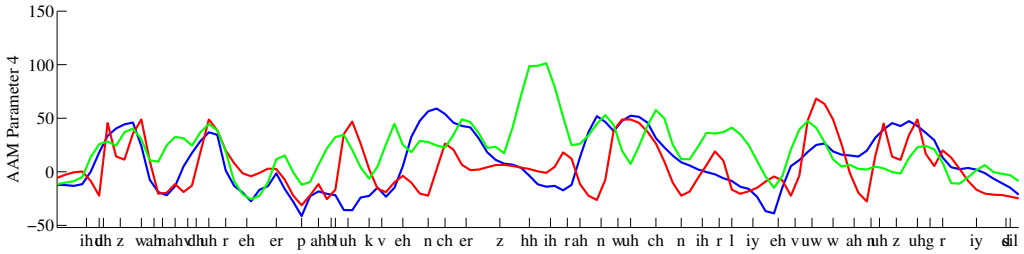
(a) Sentence 133, Parameter 1



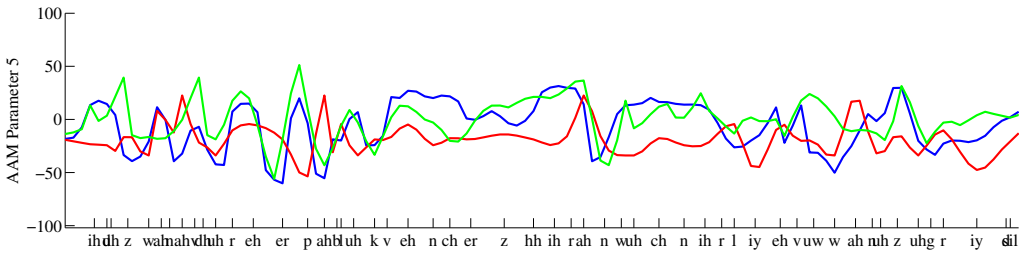
(b) Sentence 133, Parameter 2



(c) Sentence 133, Parameter 3

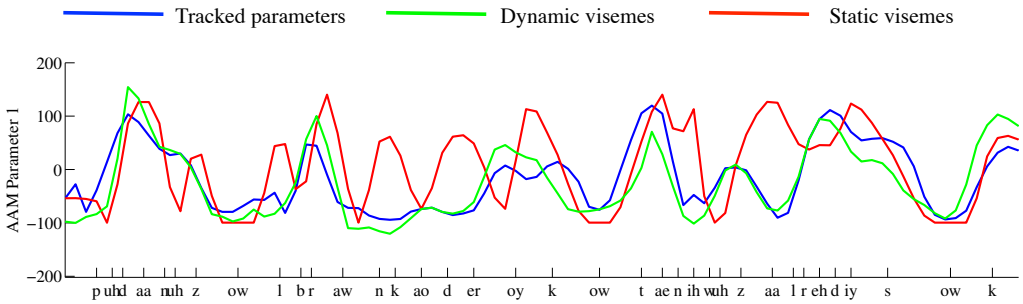


(d) Sentence 133, Parameter 4

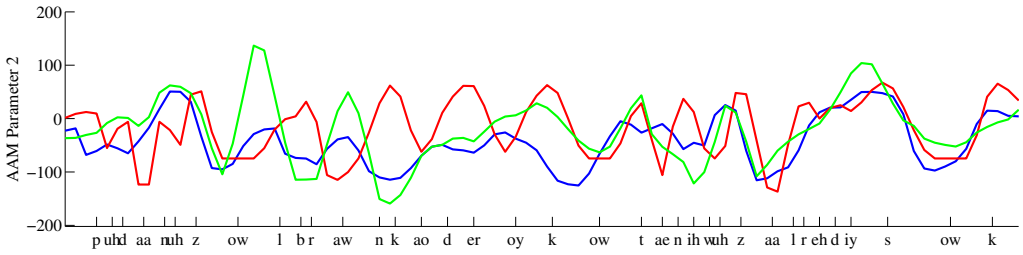


(e) Sentence 133, Parameter 5

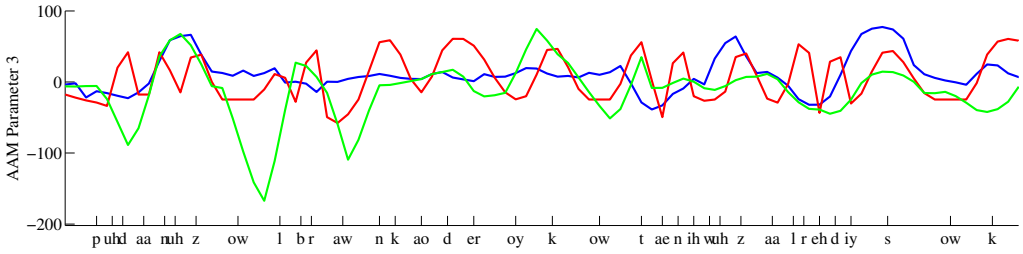
Figure C.3: The temporal trajectories for the first five combined modes of variation of a multi-segment AAM for the sentence "It is one of the rare public ventures here on which nearly everyone is agreed".



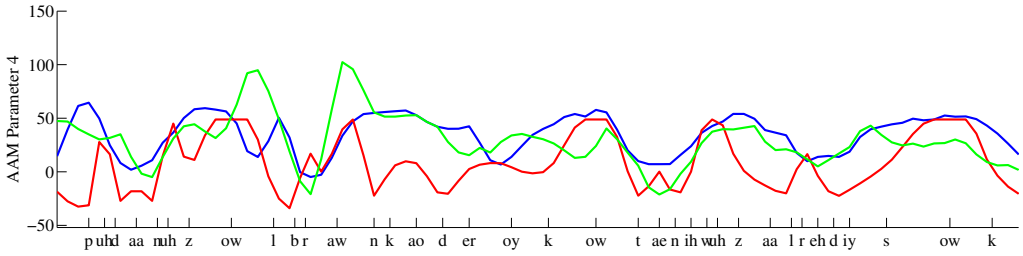
(a) Sentence 142, Parameter 1



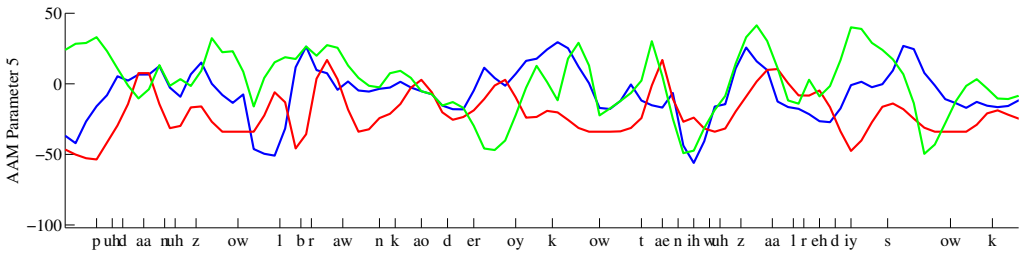
(b) Sentence 142, Parameter 2



(c) Sentence 142, Parameter 3



(d) Sentence 142, Parameter 4



(e) Sentence 142, Parameter 5

Figure C.4: The temporal trajectories for the first five combined modes of variation of a multi-segment AAM for the sentence "Put on his old brown corduroy coat and it was already soaked".

C.2 Animation Frames

This section contains frames from animated sentences which have been generated using either known dynamic viseme concatenation or static pose interpolation as described in Section 6.5.1. All of the sentences in this section are from the training data.

Figure C.5: The odd frames from an animated sequence generated using dynamic visemes for the sentence "The staff deserves a lot of credit working down here under real obstacles".

Figure C.6: The odd frames from an animated sequence generated using static pose interpolation for the sentence "The staff deserves a lot of credit working down here under real obstacles".

Figure C.7: The frames from an animated sequence generated using dynamic visemes for the sentence "Don't plan meals that are too complicated".

Figure C.8: The frames from an animated sequence generated using static pose interpolation for the sentence "Don't plan meals that are too complicated".

Figure C.9: The odd frames from an animated sequence generated using dynamic visemes for the sentence "A cardboard pattern cut to fit inside holder will help to prevent warping".

Figure C.10: The odd frames from an animated sequence generated using static pose interpolation for the sentence "A cardboard pattern cut to fit inside holder will help to prevent warping".

Figure C.11: The frames from an animated sequence generated using dynamic visemes for the sentence "The fear of punishment just didn't bother him".

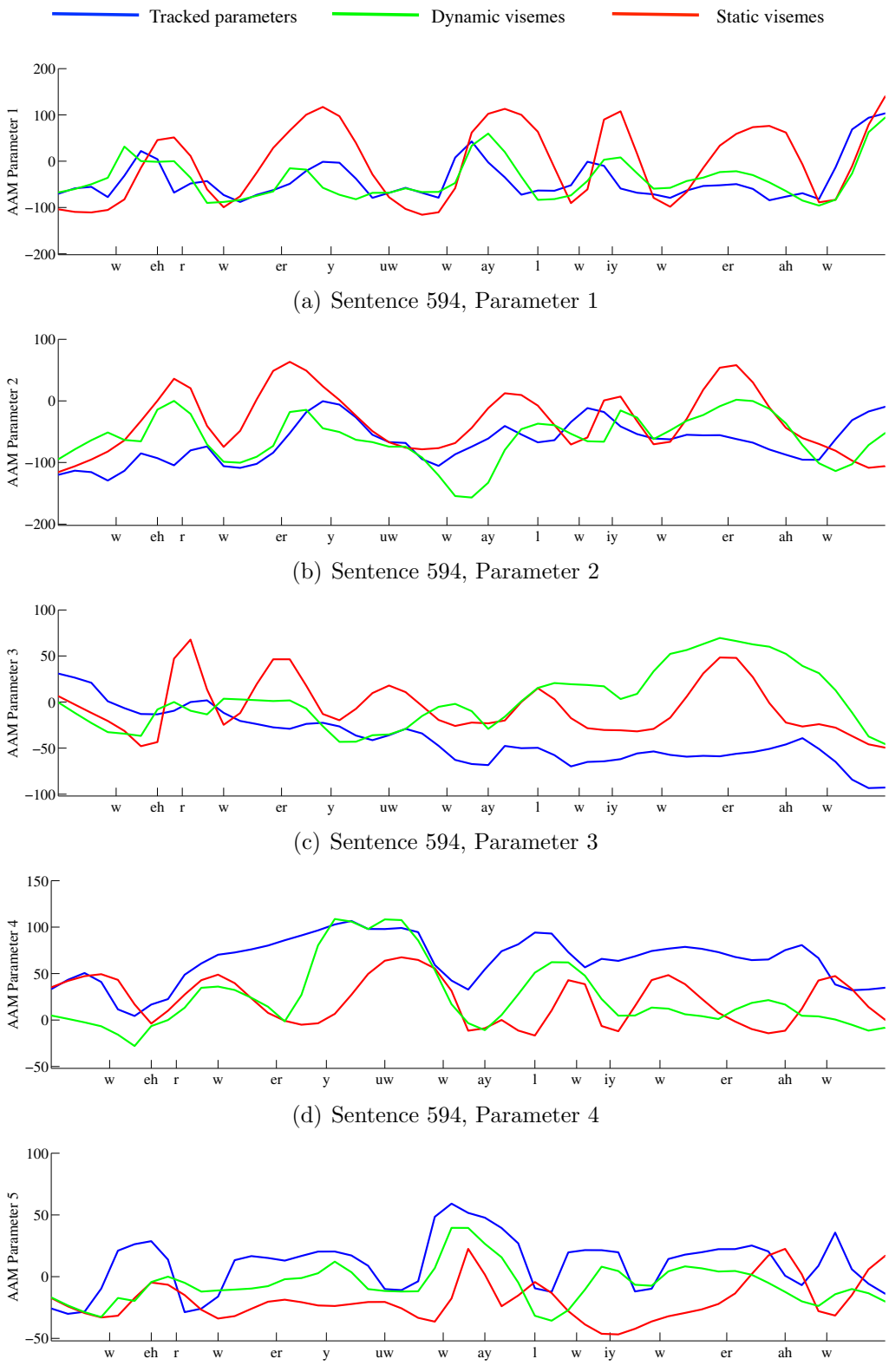
Figure C.12: The frames from an animated sequence generated using static pose interpolation for the sentence "The fear of punishment just didn't bother him".

Appendix D

Animation Output for Test Sentences

D.1 Trajectories

This section contains a selection of synthesised AAM trajectories for previously unseen sentences using the phoneme-to-dynamic viseme lookup described in Section 7.2. The generated viseme sequences are concatenated with blending at the boundary frames. For comparison, the true AAM parameters and the trajectories formed from a traditional phoneme-to-static viseme mapping are presented.



(a) Sentence 594, Parameter 1

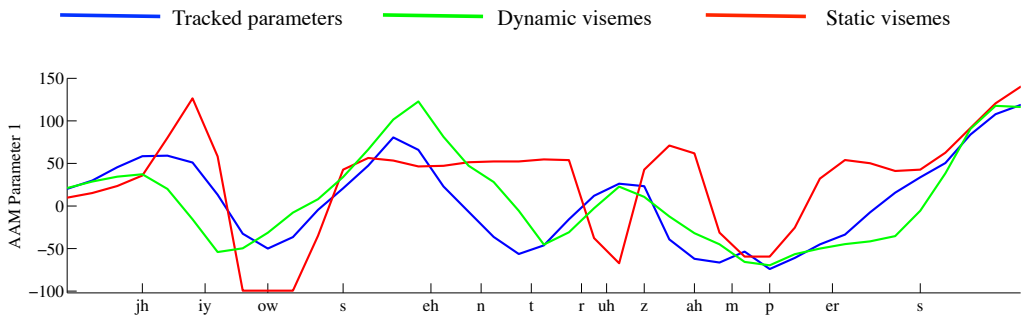
(b) Sentence 594, Parameter 2

(c) Sentence 594, Parameter 3

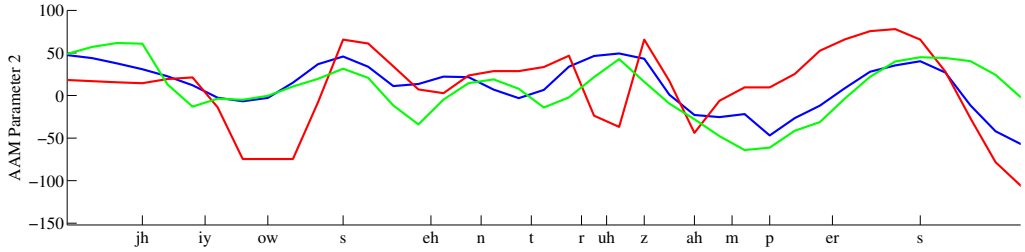
(d) Sentence 594, Parameter 4

(e) Sentence 594, Parameter 5

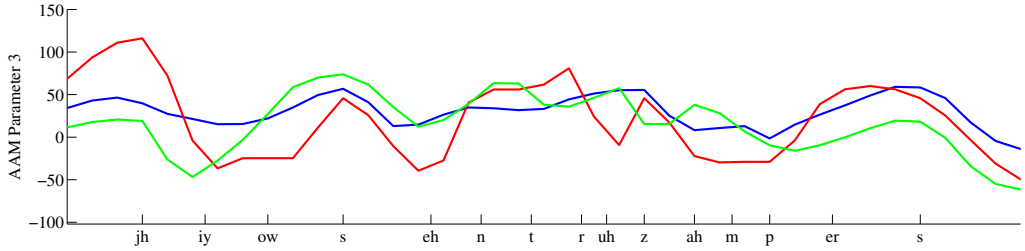
Figure D.1: The temporal trajectories for the first five combined modes of variation of a multi-segment AAM for the sentence "Where were you while we were away?". Shown are the ground-truth parameter values (blue), the concatenated dynamic viseme cluster medians for the synthesised sequences (green) and the interpolated static visemes (red).



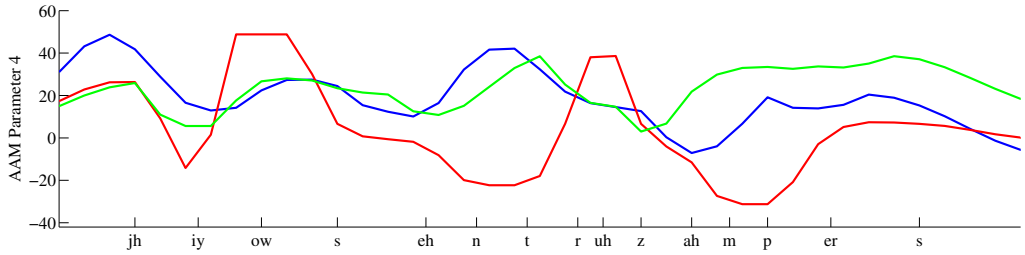
(a) Sentence 881, Parameter 1



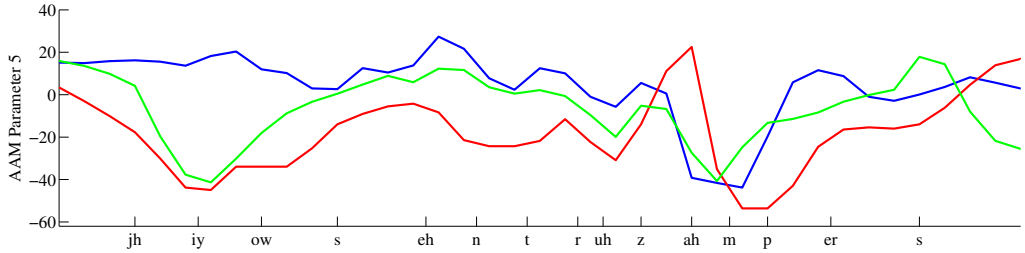
(b) Sentence 881, Parameter 2



(c) Sentence 881, Parameter 3

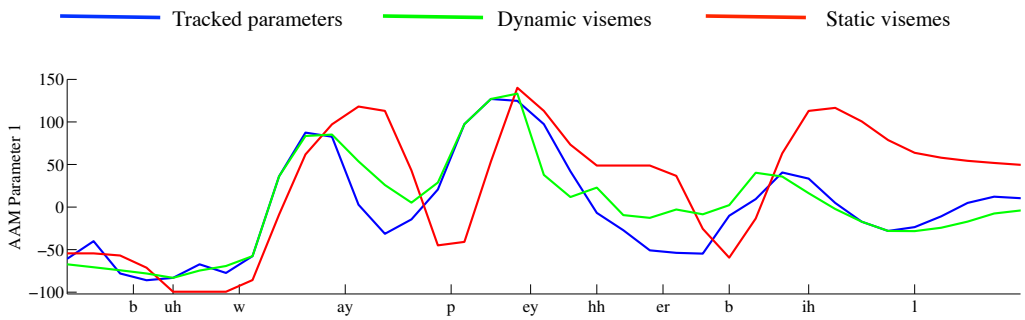


(d) Sentence 881, Parameter 4

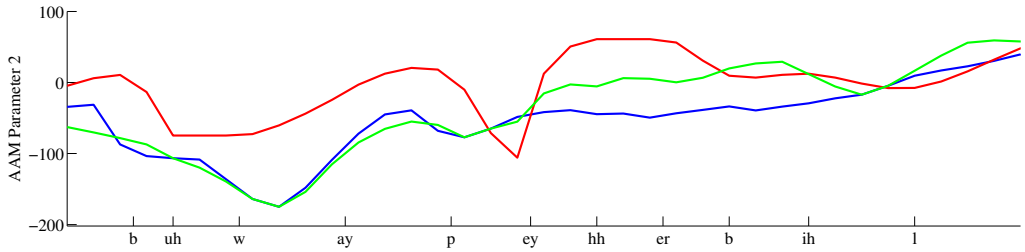


(e) Sentence 881, Parameter 5

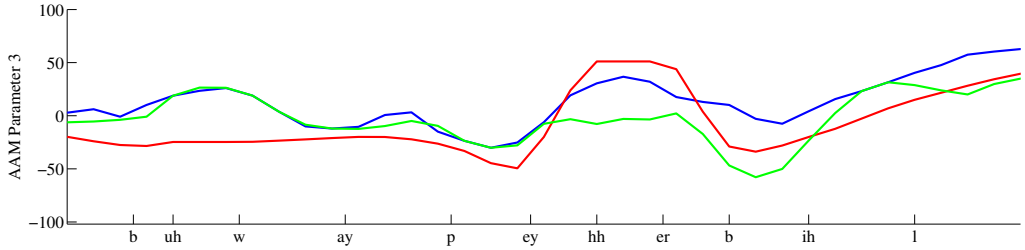
Figure D.2: The temporal trajectories for the first five combined modes of variation of a multi-segment AAM for the sentence “Geocentricism per se?”.



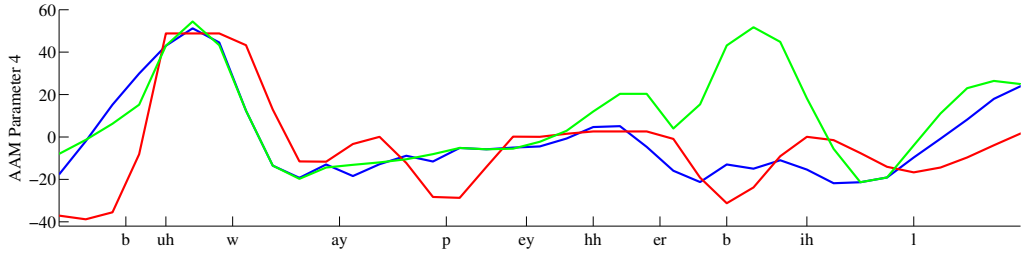
(a) Sentence 945, Parameter 1



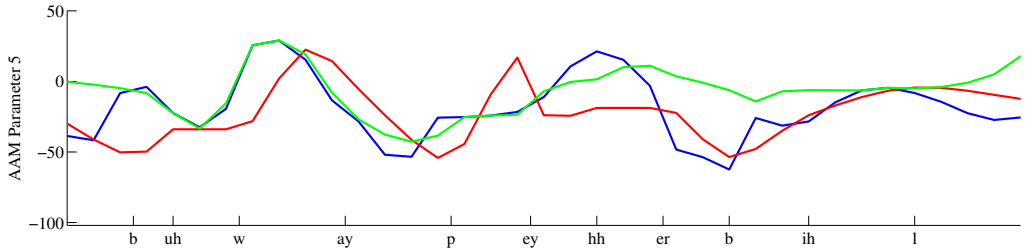
(b) Sentence 945, Parameter 2



(c) Sentence 945, Parameter 3

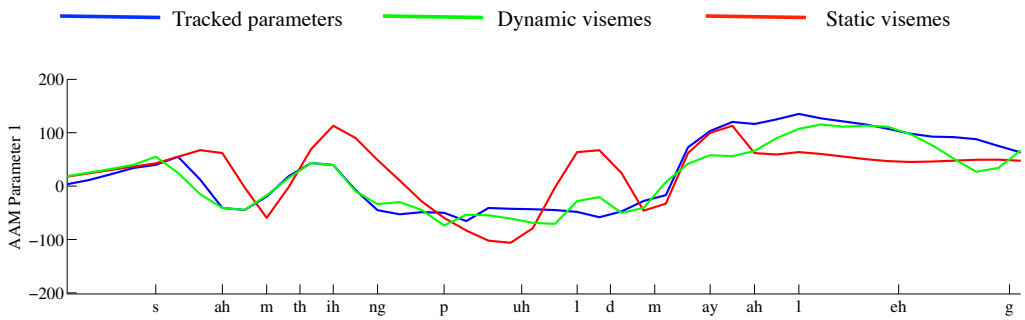


(d) Sentence 945, Parameter 4

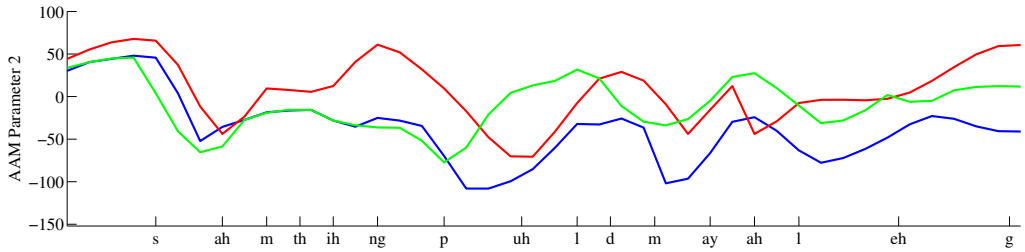


(e) Sentence 945, Parameter 5

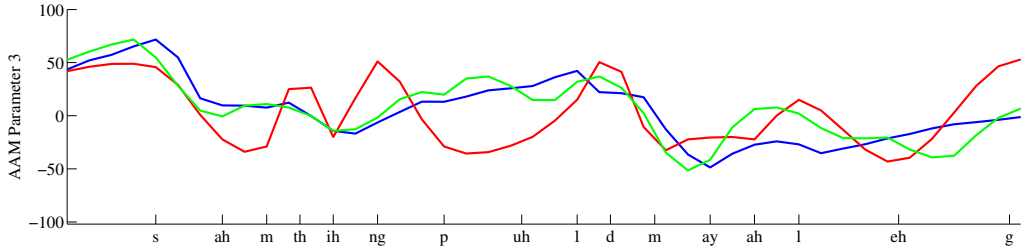
Figure D.3: The temporal trajectories for the first five combined modes of variation of a multi-segment AAM for the sentence "But why pay her bills?".



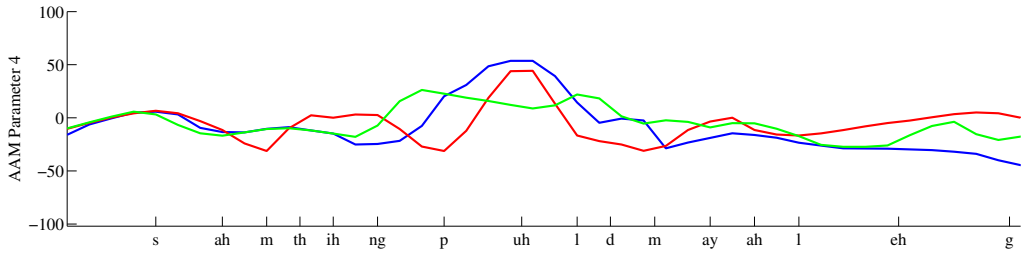
(a) Sentence 974, Parameter 1



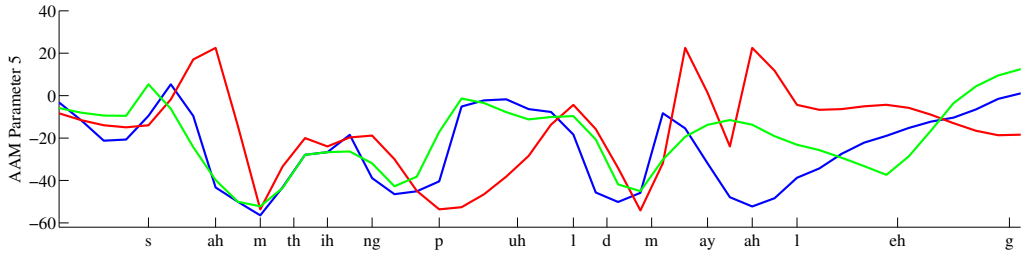
(b) Sentence 974, Parameter 2



(c) Sentence 974, Parameter 3



(d) Sentence 974, Parameter 4



(e) Sentence 974, Parameter 5

Figure D.4: The temporal trajectories for the first five combined modes of variation of a multi-segment AAM for the sentence "Something pulled my leg".

D.2 Animation Frames

This section contains animation frames for a selection of test sentences which were held out of clustering. This animations have been generated by either concentating the sequence of dynamic visemes which was determined by the phoneme to dynamic viseme mapping described in Section 7.2 or by interpolating between static poses.

Figure D.5: The odd frames from an animated sequence generated using dynamic visemes for the sentence "What obsessions had she picked up during these long nights of talk?".

Figure D.6: The odd frames from an animated sequence generated using static pose interpolation for the sentence "What obsessions had she picked up during these long nights of talk?".

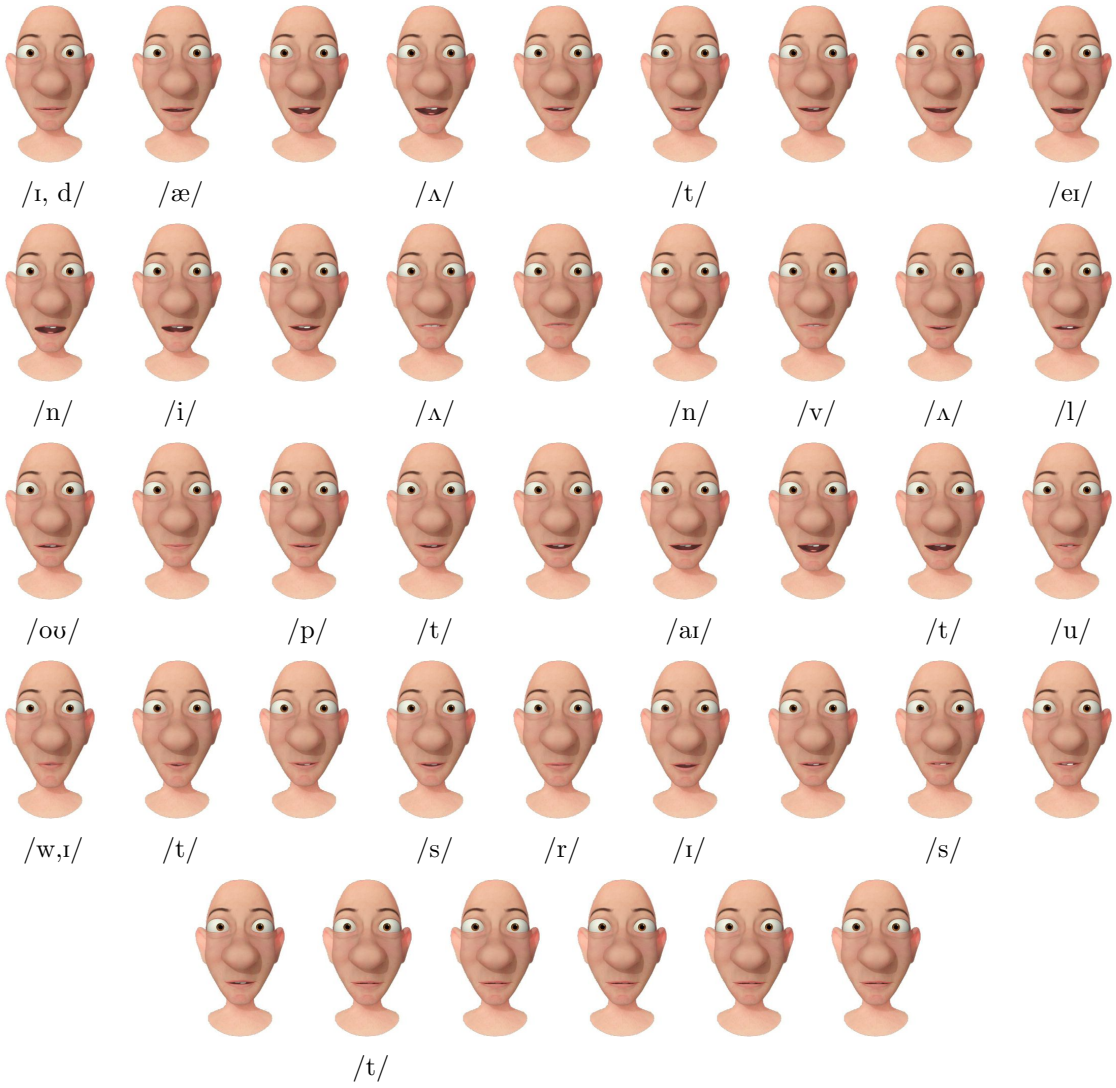


Figure D.7: The odd frames from an animated sequence generated using dynamic visemes for the sentence "It had a tiny envelope tied to its wrist".

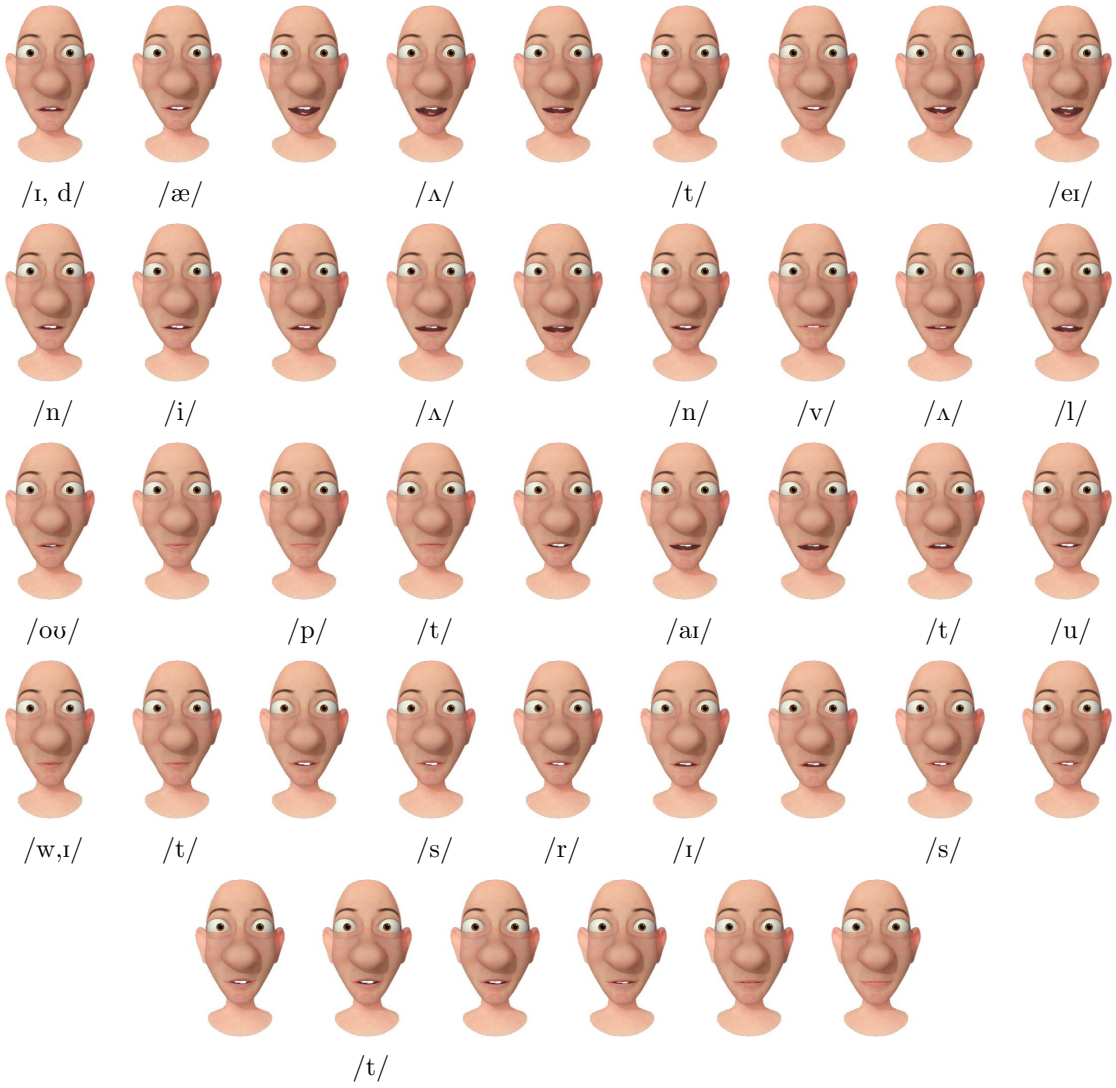


Figure D.8: The odd frames from an animated sequence generated using static pose interpolation for the sentence "It had a tiny envelope tied to its wrist".

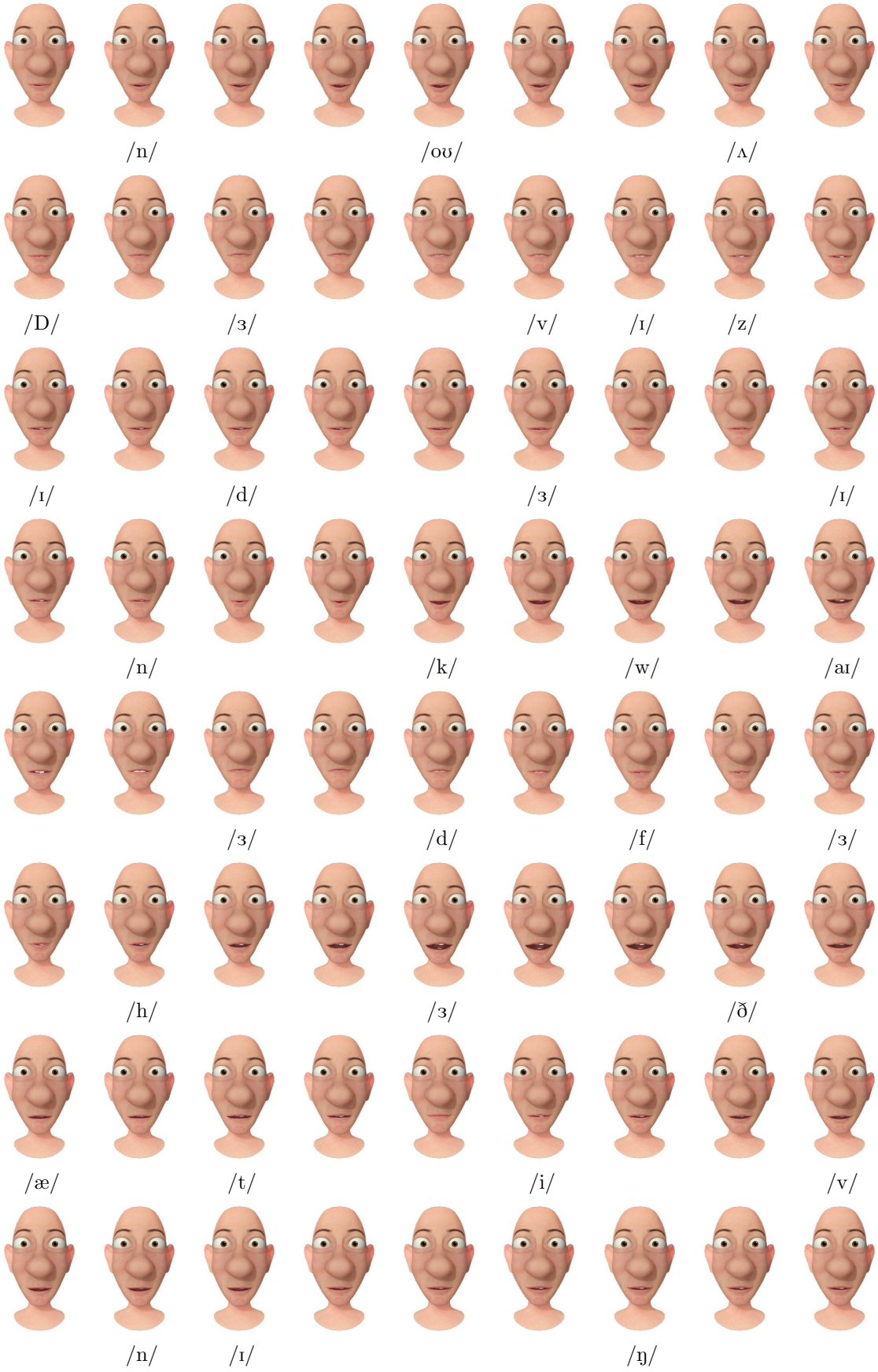


Figure D.9: The frames from an animated sequence generated using dynamic visemes for the sentence "No other visitor enquired for her that evening".

Figure D.10: The frames from an animated sequence generated using static pose interpolation for the sentence "No other visitor enquired for her that evening".

Figure D.11: The frames from an animated sequence generated using dynamic visemes for the sentence "Resistance thermometers".

Figure D.12: The frames from an animated sequence generated using static pose interpolation for the sentence “Resistance thermometers”.

Appendix E

List of Abbreviations

Abbreviation	Meaning
AAM	Active Appearance Model
ASM	Active Shape Model
C	Consonant
CCA	Canonical Correlation Analysis
CLUTO	Clustering Toolkit [77]
DCT	Discrete Cosine Transformation
DTW	Dynamic Time Warp
EMA	Electro-Magnetic Articulograph
GMM	Gaussian Mixture Model
HD	High Definition
HMM	Hidden Markov Model
HTK	Hidden Markov Model Toolkit [159]
IPA	International Phonetic Alphabet
LDA	Linear Discriminant Analysis
PCA	Principal Component Analysis
PDF	Probability Density Function
PDM	Point Distribution Model
RMS	Root Mean Squared
ROI	Region of Interest
SNR	Signal to Noise Ratio
UBM	Universal Background Model
V	Vowel

Bibliography

[1] P. Arnold and F. Hill. Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology*, 92(2):339–355, may 2001.

[2] E. Aronson and S. Rosenbloom. Space perception in early infancy: Perception within a common auditory-visual space. *Science*, 172(3988):1161–1163, 1971.

[3] E. T. Auer and L. E. Bernstein. Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness. *Journal of the Acoustical Society of America (JASA)*, 102:3704–3710, Dec. 1997.

[4] P. S. Beddor, J. D. Harnsberger, and S. Lindemann. Language-specific patterns of vowel-to-vowel coarticulation: Acoustic structures and their perceptual correlates. *Journal of Phonetics*, pages 591–627, 2002.

[5] F. Bell-Berti and K. S. Harris. Anticipatory coarticulation: Some implications from a study of lip rounding. *Journal of the Acoustical Society of America (JASA)*, 65(5):121–125, May 1979.

[6] F. Bell-Berti and K. S. Harris. A temporal model of speech production. *Phonetica*, 38(1–3):9–20, 1981.

[7] A. P. Benguerel and H. A. Cowan. Coarticulation of upper lip protrusion in french. *Phonetica*, 30:41–55, 1974.

[8] A. P. Benguerel and M. K. Pichora-Fuller. Coarticulation effects in lipreading. *Journal of Speech and Hearing Research (JSHR)*, 25:600–607, 1982.

[9] F. Bettinger, T. F. Cootes, and C. J. Taylor. Modelling facial behaviours. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG), Seoul, Korea, May 17*, pages 797–806, 2002.

[10] S. B. Bhise and A. V. Yadav. *Human Anatomy And Physiology*. Nirali Prakashan, 2008.

[11] C. A. Binnie, P. L. Jackson, and A. A. Montgomery. Visual intelligibility of consonants: A lipreading screening test with implications for aural rehabilitation. *Journal of Speech and Hearing Disorders*, 41:530–539, 1976.

[12] G. Biswas, K. Leelawong, D. Schwartz, N. Vye, and T. T. A. G. at Vanderbilt. Learning by teaching: A new agent paradigm for educational software. *Applied Artificial Intelligence*, 19(3–4):363–392, 2005.

[13] M. Brand. Voice puppetry. In *Proceedings of SIGGRAPH*, pages 21–28, Los Angeles, California, 1999.

[14] H. Bredin and G. Chollet. Measuring audio and visual speech synchrony: Methods and applications. In *IET International Conference on Visual Information Engineering (VIE)*, pages 255–260, September 2006.

[15] C. Bregler, M. Covell, and M. Slaney. Video rewrite: Driving visual speech with audio. In *Proceedings of SIGGRAPH*, pages 353–360, 1997.

[16] C. Bregler, H. Hild, S. Manke, and A. Waibel. Improving connected letter recognition by lipreading. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 557–560, April 1993.

[17] C. Bregler and Y. Konig. "Eigenlips" for robust speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 669–672, April 1994.

[18] C. Bregler and S. M. Omohundro. Learning visual models for lipreading. In M. Shah and R. Jain, editors, *Motion-based Recognition, Computational Imaging and Vision*, pages 301–320. Kluwer Academic, 1997.

[19] C. Bregler, G. Williams, S. Rosenthal, and I. McDowall. Improving acoustic speaker verification with visual body-language features. In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1909–1912, April 2009.

[20] N. Brooke and P. Templeton. Classification of lip-shapes and their association with acoustic speech events. In *The ESCA Workshop on Speech Synthesis*, pages 245–248, September 1990.

[21] E. Caldognetto, C. Zmarich, P. Cosi, and F. Ferrero. Italian consonantal visemes: Relationships between spatial/ temporal articulatory characteristics and coproduced acoustic signal. In *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP)*, pages 5–8, September 1997.

[22] R. Campbell. The processing of audio-visual speech: Empirical and neural bases. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1493):1001–1010, march 2008.

[23] W. Campbell, D. Sturim, and D. Reynolds. Support vector machines using GMM supervectors for speaker verification. *Signal Processing Letters, IEEE*, 13(5):308 – 311, 2006.

[24] Y. Cao, E. Faloutsos, P. Kohler, and F. Pighin. Real-time speech motion synthesis from recorded motions. In *Symposium on Computer Animation*, pages 347–355, 2004.

[25] L. Cappelletta and N. Harte. Viseme definitions comparison for visual-only speech recognition. In *European Signal Processing Conference (EUSIPCO)*, pages 2109–2113, Barcelona, Spain, August – September 2011.

[26] E. Chuang and C. Bregler. Mood swings: Expressive speech animation. *ACM Transactions on Graphics*, 24(2):331–347, 2005.

[27] M. M. Cohen and D. W. Massaro. Modeling coarticulation in synthetic visual speech. In *Models and Techniques in Computer Animation*, pages 139–156. Springer-Verlag, 1993.

[28] R. Cole, D. W. Massaro, J. D. Villiers, B. Rundle, K. Shobaki, J. Wouters, M. Cohen, J. Beskow, P. Stone, P. Connors, and D. Solcher. New tools for interactive speech and language training: Using animated conversational agents in the classrooms of profoundly deaf children. Presented at the ESCA/ SOCRATES Workshop on Method and Tool Innovations for Speech Science Education, 1999.

[29] M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. *Journal of the Acoustical Society of America*, 120(5):2421–2424, November 2006.

[30] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(6):681–685, June 2001.

[31] T. Cootes and C. Taylor. Statistical models of appearance for computer vision. Technical report, Department of Imaging Science and Biomedical Engineering, University of Manchester, Manchester, March 2004.

[32] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models — their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[33] E. Cosatto and H. Graf. Photo-realistic talking-heads from image samples. *IEEE Transactions on Multimedia*, 2(3):152–163, 2000.

[34] J. M. De Martino, L. P. Magalhães, and F. Violaro. Facial animation based on context-dependent visemes. *Journal of Computers and Graphics*, 30(6):971 – 980, 2006.

[35] S. Deena and A. Galata. Speech-driven facial animation using a shared Gaussian process latent variable model. In *Proceedings of the International Symposium on Advances in Visual Computing: Part I*, ISVC '09, pages 89–100, Berlin, Heidelberg, 2009. Springer-Verlag.

[36] S. Deena, S. Hou, and A. Galata. Visual speech synthesis by modelling coarticulation dynamics using a non-parametric switching state-space model. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 1–8, 2010.

[37] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1):1–38, 1977.

[38] P. Dey, S. Maddock, and R. Nicolson. Evaluation of a viseme-driven talking head. In *Proceedings of EG UK Theory and Practice of Computer Graphics*, pages 139–142, 2010.

[39] C. Dong, Y. Dong, J. Li, and H. Wang. Support vector machines based text dependent speaker verification using HMM supervectors. In *Proceedings of Speaker Odyssey: The Speaker and Language Recognition Workshop*, Stellenbosch, South Africa, January 2008.

[40] P. Duchnowski, U. Meier, and A. Waibel. See me, hear me: Integrating automatic speech recognition and lip-reading. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 547–550, 1994.

[41] P. Ekman and W. Friesen. *Facial Action Coding System*. Consulting Psychologists Press, Stanford University, Palo Alto, 1977.

[42] G. Englebienne, T. Cootes, and M. Rattray. A probabilistic model for generating realistic speech movements from speech. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1–8, December 2007.

[43] C. Engström. *Articulatory analysis of Swedish visemes*. PhD thesis, KTH, Stockholm, 2003.

[44] B. Everitt, S. Landau, and M. Leese. *Cluster Analysis*. Arnold Publishers, fourth edition, 2001.

[45] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. In *Proceedings of SIGGRAPH*, pages 388–398, 2002.

[46] T. Ezzat and T. Poggio. Miketalk: A talking facial display based on morphing visemes. In *Proceedings of the Computer Animation Conference*, pages 96–103, 1998.

[47] T. Ezzat and T. Poggio. Visual speech synthesis by morphing visemes. *International Journal of Computer Vision (IJCV)*, 38(1):45–57, 2000.

[48] M. Fehrenbach and S. Herring. *Illustrated Anatomy of the Head And Neck*. Saunders Elsevier, 2007.

[49] C. Fisher. Confusions among visually perceived consonants. *Journal of Speech and Hearing Research (JSHR)*, 11:796–804, 1968.

[50] G. Forney Jr. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

[51] C. A. Fowler. Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, 8:113–133, 1980.

[52] C. A. Fowler. Production and perception of coarticulation among stressed and unstressed vowels. *Journal of Speech and Hearing Research (JSHR)*, 24:127–139, March 1981.

[53] C. A. Fowler and E. Saltzman. Coordination and coarticulation in speech production. *Language and Speech*, 36(2, 3):171–195, 1993.

[54] W. N. Francis and H. Kucera. The Brown corpus, 1967. Brown University, Providence, RI.

[55] J. Franks and J. Kimble. The confusion of english consonant clusters in lipreading. *Journal of Speech and Hearing Research (JSHR)*, 15:474–482, September 1972.

[56] V. A. Fromkin. The non-anomalous nature of anomalous utterances. *Language*, 47(1):27–52, 1971.

[57] R. Goecke and B. Millar. Statistical analysis of the relationship between audio and video speech parameters for australian english. In *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP)*, 2003.

[58] A. J. Goldschen, O. N. Garcia, and E. Petajan. Continuous optical automatic speech recognition by lipreading. In *Proceedings of the 28th Asilomar Conference on Signals, Systems, and Computers*, pages 572–577, 1994.

[59] J. Gower. Generalized procrustes analysis. *Psychometrika*, 40:33–51, 1975.

[60] V. L. Gracco and J. H. Abbs. Variant and invariant characteristics of speech movements. *Experimental Brain Research*, 65:156–166, 1986.

[61] S. Günter and H. Bunke. Validation indices for graph clustering. *Pattern Recognition Letters*, 24(8):1107–1113, 2003.

[62] K. Harris. Coarticulation as a component in articulatory description. *Haskins Laboratories Status Report on Speech Research*, (SR-79/80):19–37, 1984.

[63] T. J. Hazen, K. Saenko, C.-H. La, and J. R. Glass. A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments. In *Proceedings of the International conference on Multimodal Interfaces (ICMI)*, pages 235–242, New York, NY, USA, 2004. ACM.

[64] W. Henke. *Dynamic articulatory model of speech production using computer simulation*. PhD thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 1966.

[65] J. Hershey and J. Movellan. Audio-vision: Using audio-visual synchrony to locate sounds. In *Advances in Neural Information Processing Systems 12*, pages 813–819. MIT Press, 2000.

[66] G. Hripcsak and A. Rothschild. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298, 2005.

[67] P. J. Hsieh, J. T. Colas, and N. Kanwisher. Spatial pattern of BOLD fMRI activation reveals cross-modal information in auditory cortex. *Journal of Neurophysiology*, 107:3428–3432, April 2012.

[68] L. S. Hultzen. Tables of transitional frequencies of english phonemes, 1964. University of Illinois Press.

[69] A. Hunt and A. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 373 – 376, May 1996.

[70] F. J. Ijsseldijk. Speechreading performance under different conditions of video image, repetition, and speech rate. *Journal of Speech and Hearing Research (JSHR)*, 35(2):466, 1992.

[71] International Phonetic Association. *Handbook of the International Phonetic Association, A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, July 1999.

[72] P. J. Jackson and V. D. Singampalli. Statistical identification of articulation constraints in the production of speech. *Speech Communication*, 51(8):695–710, 2009.

[73] P. L. Jackson. The theoretical minimal unit for visual speech perception: Visemes and coarticulation. *Volta Review*, 90(5):99–115, September 1988.

[74] A. Jain, M. Murty, and P. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[75] J. Jiang, A. Alwan, L. E. Bernstein, E. T. Auer, and P. A. Jr. Keating. Similarity structure in perceptual and physical measures for visual consonants across talkers. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 441–444, 2002.

[76] D. Jones. *An outline of English phonetics*. Cambridge University Press, 9th edition, April 1976.

[77] G. Karypis. *CLUTO — A clustering toolkit*. University of Minnesota, Department of Computer Science, Minneapolis, April 2002.

[78] G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20(1):359–392, December 1998.

[79] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[80] R. D. Kent and F. D. Minifie. Coarticulation in recent speech production models. *Journal of Phonetics*, 5:115–133, 1977.

[81] R. D. Kent and K. L. Moll. Tongue body articulation during vowel and diphthong gestures. *Folia Phoniatrica*, 24:278–300, 1972.

[82] E. Keogh and C. A. Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and Information Systems*, 7(3):358–386, 2005.

[83] C. Kouadio, P. Poulin, and P. Lachapelle. Real-time facial animation based upon a bank of 3D facial expressions. In *Proceedings of the Computer Animation*, CA '98, pages 128–137, Washington, DC, USA, 1998. IEEE Computer Society.

[84] V. Kozhevnikov and L. Chistovich. *Speech: articulation and perception*. English translation: Joint publication Research Service, Washington D.C., 1965.

[85] R. Krakow. Nonsegmental influences on velum movement patterns: Syllables, sentences, stress, and speaking rate. *Phonetics and Phonology: Nasals, Nasalization, and the Velum*, 5:87–116, 1993.

[86] P. Ladefoged, R. Harshman, L. Goldstein, and L. Rice. Generating vocal tract shapes from formant frequencies. *Journal of the Acoustical Society of America*, 64(4):1027–1035, October 1978.

[87] P. Ladefoged and K. Johnson. *A course in phonetics*. Wadsworth, 6th edition, 2011.

[88] Y. Lan, R. Harvey, B. Theobald, E. Ong, and R. Bowden. Comparing visual features for lipreading. In *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP)*, pages 102–106, 2009.

[89] O. Lazalde and S. Maddock. Comparison of different types of visemes using a constraint-based coarticulation model. In *Theory and Practice of Computer Graphics*, pages 199–206. The Eurographics Association, 2010.

[90] S. Lee and D. Yook. Audio-to-visual conversion using hidden Markov models. In *Proceedings of the 7th Pacific Rim International Conference on Artificial Intelligence: Trends in Artificial Intelligence (PRICAI)*, pages 563–570, London, UK, 2002. Springer-Verlag.

[91] S. Lesner and P. Kricos. Visual vowel and diphthong perception across speakers. *Journal of the Academy of Rehabilitative Audiology (JARA)*, pages 252–258, 1981.

[92] S. Lesner, S. Sandridge, and P. Kricos. Training influences on visual consonant and sentence recognition. *Ear and Hearing*, 8(5), 1987.

[93] V. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics — Doklady, Cybernetics and Control Theory*, 10(8):707–710, 1966.

[94] B. Lidestam and J. Beskow. Visual phonemic ambiguity and speechreading. *Journal of Speech, Language and Hearing Research (JSLHR)*, 49:835–847, August 2006.

[95] A. Löfqvist. Speech as audible gestures. *Speech production and speech modelling*, pages 289–322, 1990.

[96] P. Lucey. *Lipreading across Multiple Views*. PhD thesis, Queensland University of Technology, Brisbane, Queensland, 2007.

[97] P. J. Lucey, S. Sridharan, and D. B. Dean. Continuous pose-invariant lipreading. In *Proceedings of Interspeech*, pages 2679–2682, 2008.

[98] J. Luettin. *Visual Speech and Speaker Recognition*. PhD thesis, University of Sheffield, Sheffield, United Kingdom, 1997.

[99] J. Luettin and N. A. Thacker. Speechreading using probabilistic models. *Computer Vision and Image Understanding*, 65(2):163–178, 1997.

[100] J. Ma, R. Cole, B. Pellom, W. Ward, and B. Wise. Accurate automatic visible speech synthesis of arbitrary 3D models based on concatenation of diviseme motion capture data. *Journal of Computer Animation and Virtual Worlds*, 15(5):485–500, December 2004.

[101] J. Ma, R. Cole, B. Pellom, W. Ward, and B. Wise. Accurate visible speech synthesis based on concatenating variable length motion capture data. *IEEE Transactions on Visualization and Computer Graphics*, 12(2):266–276, March 2006.

[102] H. Magen. The extent of vowel-to-vowel coarticulation in english. *Journal of Phonetics*, 25(2):187–205, 1997.

[103] S. Y. Manuel. The role of contrast in limiting vowel-to-vowel coarticulation in different languages. *Journal of the Acoustical Society of America (JASA)*, pages 1–20, 1990.

[104] D. Massaro. *Perceiving talking faces: From speech perception to a behavioral principle*. The MIT Press, 1998.

[105] D. Massaro and J. Light. Improving the vocabulary of children with hearing loss. *The Volta Review*, 104(3):141–174, 2004.

[106] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision (IJCV)*, 60(2):135–164, 2004.

[107] I. Matthews, T. Cootes, A. Bangham, S. Cox, and R. Harvey. Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 24(2):198–213, february 2002.

[108] W. Mattheyses, L. Latacz, and W. Verhelst. Automatic viseme clustering for audiovisual speech synthesis. In *Proceedings of Interspeech*, pages 2173–2176, 2011.

[109] S. L. Mattys, L. E. Bernstein, and E. T. Auer. Stimulus-based lexical distinctiveness as a general word-recognition mechanism. *Perception and Psychophysics*, 64(4):667–679, 2002.

[110] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, Dec. 1976.

[111] J. Melenchón, J. Simó, G. Cobo, and E. Martínez. Objective viseme extraction and audiovisual uncertainty: Estimation limits between auditory and visual modes. In *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP)*, 2007.

[112] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The extended M2VTS database. In *Proceedings of the International Conference on Audio and Video-based Biometric Person Authentication*, pages 72–76, March 1999.

[113] G. Modarresi, H. Sussman, B. Lindblom, and E. Burlingame. An acoustic analysis of the bidirectionality of coarticulation in VCV utterances. *Journal of Phonetics*, 32(3):291 – 312, 2004.

[114] A. A. Montgomery and P. L. Jackson. Physical characteristics of the lips underlying vowel lipreading performance. *Journal of the Acoustical Society of America (JASA)*, 73(6):2134–2144, 1983.

[115] J. Navarra and S. Soto-Faraco. Hearing lips in a second language: Visual articulatory information enables the perception of second language sounds. *Psychological Research*, 71:4–12, 2007.

[116] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou. Audio-visual speech recognition, final workshop 2000 report. Technical report, Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD, October 2000.

[117] NIST. The DARPA TIMIT acoustic-phonetic continuous speech corpus (TIMIT). [CD-ROM], November 1988.

[118] J. Noh and U. Neumann. Expression cloning. In *Proceedings of SIGGRAPH*, pages 277–288, 2001.

[119] E. Owens and B. Blazek. Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech and Hearing Research (JSHR)*, 28:381–393, 1985.

[120] F. I. Parke. Computer generated animation of faces. In *Proceedings of the ACM annual conference*, pages 451–457, New York, NY, USA, 1972. ACM.

[121] F. I. Parke and K. Waters. *Computer Facial Animation*. A. K. Peters, Ltd., Natick, MA, USA, 1996.

[122] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy. Cuave: A new audio-visual database for multimodal human-computer interface research. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2017–2020, 2002.

[123] C. Pelachaud. *Communication and coarticulation in facial animation*. PhD thesis, University of Pennsylvania, Philadelphia, PA, USA, 1992.

[124] C. Pelachaud, N. Badler, and M. Steedman. Linguistics issues in facial animation. *Computer Animation*, pages 15–30, 1991.

[125] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. In *ACM SIGGRAPH*, pages 313–318, New York, NY, USA, 2003. ACM.

[126] J. S. Perkell and M. L. Matthies. Temporal measures of anticipatory labial coarticulation for the vowel /u/: Within- and cross-subject variability. *Journal of the Acoustical Society of America (JASA)*, 91(5):2911–2925, 1992.

[127] E. Petajan, B. Bischoff, and D. Bodoff. An improved automatic lipreading system to enhance speech recognition. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '88, pages 19–25, 1988.

[128] G. Potamianos and C. Neti. Improved ROI and within frame discriminant features for lipreading. In *Proceedings of the International conference in Image processing*, volume 3, pages 250–253, 2001.

[129] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[130] M. Ramage. *Disproving Visemes As The Basic Visual Unit Of Speech*. PhD thesis, School of Engineering, Department of Mechanical Engineering, Curtin University, September 2011.

[131] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1–3):19–41, 2000.

[132] P. Roach. On the distinction between 'stress-timed' and 'syllable-timed' languages. *Linguistic controversies*, pages 73–79, 1982.

[133] D. Rosenbaum. *Human motor control*. Academic Press, 2009.

[134] L. A. Ross, D. Saint-Amour, V. M. Leavitt, D. C. Javitt, and J. J. Foxe. Do you see what i am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, 17(5):1147–1153, 2007.

[135] P. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[136] J. S. Scheinberg. Analysis of speechreading cues using an interleaved technique. *Journal of Communication Disorders*, 13(6):489–492, 1980.

[137] A. Sell and M. Kaschak. Does visual speech information affect word segmentation? *Memory and Cognition*, 37(6):889–894, 2009.

[138] S. Soto-Faraco, J. Navarra, W. Weikum, A. Vouloumanos, N. Sebastián-Gallés, and J. Werker. Discriminating languages by speech-reading. *Attention, Perception and Psychophysics*, 69:218–231, 2007.

[139] W. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America (JASA)*, 26(2):212–215, 1954.

[140] Q. Summerfield. Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 335(1273):71–78, 1992.

[141] D. Terzopoulos and K. Waters. Physically-based facial modelling, analysis, and animation. *The Journal of Visualization and Computer Animation*, 1(2):73–80, 1990.

[142] D. Terzopoulos and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:569–579, 1993.

[143] B. Theobald, S. Fagel, G. Bailly, and F. Elisei. LIPS2008: Visual speech synthesis challenge. In *Proceedings of Interspeech*, pages 1875–1878, 2008.

[144] B. Theobald and N. Wilkinson. Real-time speech driven talking heads using active appearance models. In *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP)*, pages 264–269, 2007.

[145] B. J. Theobald, I. Matthews, J. R. Spies, T. R. Brick, J. F. Cohn, S. M. Boker, and M. Mangini. Mapping and manipulating facial expression. *Language and Speech*, 52(2/3):369 – 386, 2009.

[146] M. Turk and A. Pentland. Face recognition using Eigenfaces. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–591, June 1991.

[147] A. Turkmani, A. Hilton, P. Jackson, and J. Edge. Visual analysis of lip coarticulation in VCV utterences. In *Proceedings of Interspeech*, pages 1406–1409, 2007.

[148] V. van Wassenhove, K. W. Grant, and D. Poeppel. Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America*, 102(4):1181–1186, 2005.

[149] E. Vatikiotis-bateson, K. G. Munhall, M. Hirayama, Y. Lee, and D. Terzopoulos. The dynamics of audiovisual behavior in speech. In *Speechreading by Humans and Machines: Models, Systems, and Applications, volume 150 of NATO ASI Series. Series F: Computer and Systems Sciences*, pages 221–232. Springer-Verlag, 1996.

[150] D. Vlasic, M. Brand, H. Pfister, and J. Popovic. Face transfer with multilinear models. *ACM Transactions on Graphics*, 24(3):426–433, 2005.

[151] U. von Luxburg, R. Williamson, and I. Guyon. Clustering: Science or art? In *Journal of Machine Learning Research (JMLR)*, number 27 in Workshop on Unsupervised and Transfer Learning, pages 65–79, 2012.

[152] B. Walden, R. Prosek, A. Montgomery, C. Scherr, and C. Jones. Effects of training on the visual recognition of consonants. *Journal of Speech, Language and Hearing Research (JSLHR)*, 20(1):130–145, 1977.

[153] K. Wampler, D. Sasaki, L. Zhang, and Z. Popović. Dynamic, expressive speech animation from a single mesh. In *Symposium on Computer Animation*, pages 53–62, 2007.

[154] W. A. Wickelgren. Context-sensitive coding, associative memory and serial order in speech behaviour. *Psychological Review*, 76:1–15, 1969.

[155] G. Williams, G. Taylor, K. Smolskiy, and C. Bregler. Body motion analysis for multi-modal identity verification. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 2198–2201, August 2010.

[156] L. Williams. Performance driven facial animation. *Computer Graphics*, 24(2):235–242, 1990.

[157] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(13):37–52, 1987.

[158] S. Young and P. Woodland. State clustering in hidden Markov model-based continuous speech recognition. *Computer Speech and Language*, 8(4):369 – 383, 1994.

[159] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland. *The HTK Book, version 3.4*, 2006.

[160] H. Zhao and C. Tang. Visual speech synthesis based on Chinese dynamic visemes. In *Proceedings of the International Conference on Information and Automation*, pages 139 –143, June 2008.